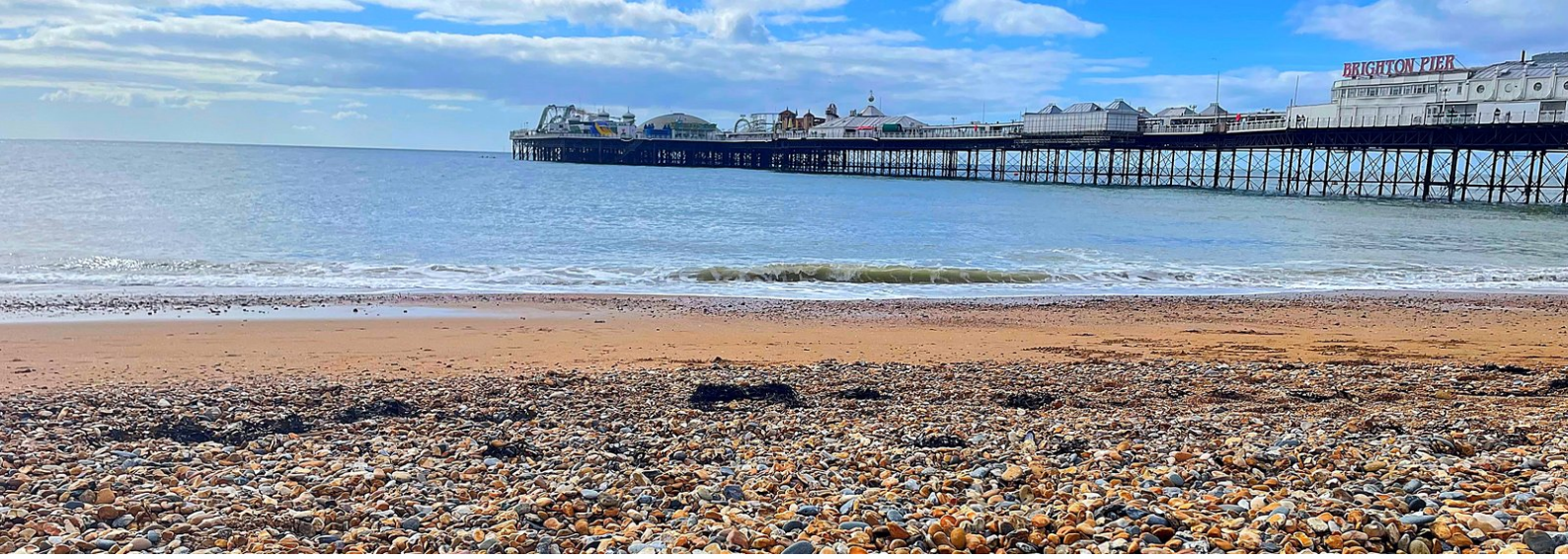


HealTAC 2026

Health Text Analytics Conference

Book of Abstracts

8–10 June 2026 · Brighton, UK

healtac2026.github.io

SPONSORS & PARTNERS

DARE UK · HDRUK Health Data Research UK · King's College London · brighton and sussex medical school · SAFETEXT · CogStack · frontiers · NIHR | Maudsley Biomedical Research Centre with NIHR Mental Health Translational Research Collaboration

Contents

Contents

Contents

Investigating Bias in Mental Health Clinical Notes

Shubham Agarwal[*1], Jaya Chaturvedi[*1], Julia Ive[*1,3], Robert Stewart[2], Thomas Searle[1], Richard Dobson[1,3]

[1]Department of Biostatistics & Health Informatics, King's College London, UK
[2]Department of Psychological Medicine, King's College London, UK
[3]University College London, UK

1 Introduction

Clinical notes are central to mental health care. In psychiatry especially, where diagnosis and care planning rely heavily on narrative interpretation, clinical documentation plays a decisive role in shaping patient trajectories. However, clinical notes may also reflect implicit biases held by healthcare providers or embedded within institutional practices. Subtle linguistic framing, differential documentation detail, or assumptions about patient behavior may contribute to disparities in diagnoses, treatment decisions, and patient outcomes [1, 2].

Prior research has demonstrated racial bias in Electronic Health Records (EHR) documentation, including disproportionate use of negative descriptors for Black patients [3] and systematic linguistic differences associated with gender and ethnicity [4]. Longstanding evidence also highlights racial and ethnic disparities in emergency and psychiatric care delivery [5]. Similarly, individuals diagnosed with Severe Mental Illness (SMI) may receive documentation that emphasizes chronicity, dangerousness, or reduced capacity in ways that differ systematically from non-SMI patients [6].

Beyond documentation practices, psychiatric diagnostic processes themselves may reflect structural and implicit bias. Multiple studies have shown that Black patients in the United States and the United Kingdom are disproportionately diagnosed with schizophrenia and other psychotic disorders compared to White patients, even when presenting with comparable affective symptoms [7, 8]. Such findings raise important questions about whether diagnostic disparities reflect true differences in prevalence or differential interpretation and labeling of similar symptom presentations.

Accordingly, this project hypothesizes that disparities may be observable not only in linguistic framing but also in diagnostic distributions and diagnostic justification within mental health notes. Specifically, we will examine whether certain demographic groups are more likely to receive SMI diagnoses, or particular diagnostic labels, after accounting for documented symptoms and clinical severity indicators. This project therefore aims to systematically identify and

[*]These authors contributed equally.

characterize implicit biases present in unstructured mental health clinical notes within the Clinical Record Interactive Search (CRIS) database [9], examining both documentation practices and diagnostic patterns across gender, race, and SMI status.

2 Methods and Data

We will analyze unstructured clinical notes drawn from the CRIS database, a repository of de-identified mental health records from the UK National Health Service. The study cohort comprises patients with their first-ever accepted referral to SLAM in 2024 who received a primary SMI diagnosis during that referral, defined by ICD-10 codes F2* (schizophrenia spectrum disorders), F30 (manic episode), or F31 (bipolar affective disorder). Patient records are stratified by gender, SMI diagnosis, and race/ethnicity.
Our approach combines natural language processing (NLP) and bias-probing prediction modeling:

1. **Feature Extraction**:
    - Linguistic tone and sentiment: evaluative descriptors, affective language, risk-focused framing
    - Negative descriptors: e.g., *non-compliant*, *guarded*, *dramatic*, *manipulative*
    - Topic and phrase distributions to detect recurring themes across demographic and diagnostic groups
2. **Bias-Probing Prediction Modeling**:
    - A classifier is trained to predict SMI diagnosis
    - The same note is tested under different demographic labels. If predicted outcomes differ (e.g., SMI for a Black patient vs. non-SMI for a White patient with identical notes), this reveals systematic bias influencing diagnostic perception
3. **Analysis**:
    - Comparison of linguistic features, note length, and negative descriptors across demographic groups
    - Intersectional analysis of gender $\times$ race $\times$ SMI status to detect compound bias
    - Adjustment for confounders including documented symptom severity, age, and care setting

3 Results

Preliminary qualitative assessment demonstrates that our methods enable effective identification and contextualization of potential biases in clinical notes. The combination of NLP-based feature extraction and bias-probing predictive modeling has the potential to provide insights into

how language and documentation patterns vary across demographic and diagnostic groups. Intersectional analysis allows for detection of compound biases, and continued work will quantify these patterns using statistical metrics of prevalence and impact.

4 Conclusion

This study aims to systematically investigate implicit bias in mental health clinical notes, focusing on both linguistic framing and diagnostic interpretation. By combining textual analysis with bias-probing predictive modeling, it provides a framework for quantifying how documentation patterns may differ across gender, race, and SMI status. The anticipated outcomes of this work include supporting clinician awareness and training programs addressing implicit bias and guiding the development of NLP tools to monitor and mitigate biased language.

5 Study context

This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and Kings College London. No conflict of interest declared.

References

[1] Hall WJ, Chapman MV, Lee KM, et al. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: A systematic review. American Journal of Public Health. 2015;105(12):e60-76.

[2] Chapman EN, Kaatz A, Carnes M. Physicians and implicit bias: How doctors may unwittingly perpetuate health care disparities. Journal of General Internal Medicine. 2013;28(11):1504-10.

[3] Sun M, Oliwa T, Peek ME, Tung EL. Negative Patient Descriptors: Documenting Racial Bias in the Electronic Health Record. Health Affairs. 2022;41(2):203-11.

[4] Markowitz DM. Gender and ethnicity bias in medicine: A text analysis of 1.8 million critical care records. PNAS Nexus. 2022;1(4):pgac157.

[5] Richardson LD, Babcock Irvin C, Tamayo-Sarver JH. Racial and ethnic disparities in the clinical practice of emergency medicine. Academic Emergency Medicine. 2003;10(11):1184-8.

[6] Schwartz RC, Blankenship DM. Racial disparities in psychotic disorder diagnosis: A review of empirical literature. World Journal of Psychiatry. 2014;4(4):133-40.

[7] Olbert CM, Nagendra A, Buck B. Meta-analysis of Black vs. White racial disparity in schizophrenia diagnosis in the United States: Do structured assessments attenuate racial disparities? Journal of Abnormal Psychology. 2018;127(1):104-15.

[8] Barnett P, Mackay E, Matthews H, et al. Ethnic variations in compulsory detention and diagnosis of psychosis: A meta-analysis. The Lancet Psychiatry. 2019;6(4):305-17.

[9] Stewart R, Soremekun M, Perera G, Broadbent M, Callard F, Denis M, et al. The South London and Maudsley NHS foundation trust biomedical research centre (SLAM BRC) case register: development and descriptive data. BMC psychiatry. 2009;9(1):51.

GS-BrainText: A Multi-Site Brain Imaging Report Dataset for Clinical NLP Development and Validation

Beatrice Alex[1,2,3], Claire Grover[4], Arlene Casey[5], Richard Tobin[4], Heather Whaley [6,7,8], William Whiteley[6,7,9]

[1] Advanced Care Research Centre, University of Edinburgh (UoE), Edinburgh, UK
[2] Edinburgh Futures Institute, UoE, Edinburgh, UK
[3] School of Literatures, Languages and Cultures, UoE, Edinburgh, UK
[4] School of Informatics, UoE, Edinburgh, UK
[5] Usher Institute, UoE, Edinburgh, UK
[6] Centre for Clinical Brain Sciences, UoE, Edinburgh, UK
[7] Institute for Neuroscience and Cardiovascular Research, UoE, Edinburgh, UK
[8] Generation Scotland, Institute of Genetics and Cancer, UoE, Edinburgh, UK
[9] NHS Lothian, Edinburgh, UK

Introduction

Clinical natural language processing (NLP) can unlock valuable information from unstructured electronic health records, but translation to clinical practice remains challenging due to limited access to diverse, representative and expertly annotated datasets. Most publicly available clinical text datasets originate from US healthcare systems and/or single institutions (e.g. [1,2]), limiting their utility for developing robust and generalisable NLP systems.

We present **GS-BrainText,** a curated dataset of 8,511 brain radiology reports from the Generation Scotland cohort, of which 2,431 are expertly annotated for 24 brain disease phenotypes. This multi-site dataset spans five Scottish NHS health boards and includes broad age representation (mean age 58, median age 53), making it valuable for developing and evaluating generalisable clinical NLP algorithms.

Generation Scotland is a population-based health study of approximately 24,000 participants [3-5], established to investigate genetic and environmental contributions to common diseases, with participants consenting to health record linkage, including radiology reports across multiple NHS health boards.[1] GS-BrainText draws on brain imaging reports (CT and MRI) for this cohort, authored by consultant radiologists from 1994 to 2021 and ranging from a few to 602 words (mean: 82 words).

Methods and Data

Data Collection and Curation: Reports were acquired via established data linkage protocols from NHS Fife, Lothian, Greater Glasgow and Clyde (GGC), Grampian and Tayside. Our in-house NLP pipeline, EdIE-R, converted raw CSV data into a structured XML format, automatically identifying sections such as clinical history and report body [6,7]. The final corpus consists of 8,511 reports (4,362 CT; 3,966 MRI).

Annotation Schema: Expert annotations created for a subset of 2,432 reports (CT: 1,487, MRI: 944) for scans spanning 1994-2021 by a multidisciplinary clinical team using a comprehensive annotation schema developed with neurologists and radiologists and using

[1] Generation Scotland now also includes an additional 17,000 participants from across the whole of Scotland (NextGenScot), including participants who are younger and from more health boards. Their data is not included in GS-BrainText.

the BRAT tool [8]. This schema comprises text-level annotations (named entities, location and temporal modifiers), attributes (negation), relations and document-level phenotype labels. The 24 phenotypes include stroke-related labels (covering ischaemic and haemorrhagic stroke subtypes by location and timing), tumour-related labels (including meningioma, metastasis, glioma) and other neurological findings (including small vessel disease, atrophy, subdural haematoma, subarachnoid haemorrhage, microbleeds and haemorrhagic transformation).

Quality Assurance (QA): Rigorous QA included 10-100% double-annotation (depending on the health board) and regular consensus meetings. Inter-annotator agreement (IAA) reached F1-scores of 83.74 to 95.65, indicating high reliability in the gold standard labels.

Results

The phenotype distribution reflects population-based epidemiology, with small vessel disease (n=545) and atrophy (n=441) being most common and other clinically important phenotypes remaining (microbleeds n=12, haemorrhagic transformation n=10, gliomas n=8). The dataset shows substantial variation in phenotype frequencies across health boards, reflecting different patient populations and clinical practices.

We evaluated the EdIE-R system (a rule-based NLP pipeline) against the GS-BrainText document-level gold standard annotation to establish a performance baseline as a starting point for future benchmarking using state-of-the-art NLP models. Components of EdIE-R have been previously evaluated on stroke register and routine brain imaging data [9,10], providing confidence in its suitability as a baseline for this task.

EdIE-R performance achieved 88.82 micro-averaged F1 and > 95 F1 for high-frequency phenotypes, though it dropped significantly for rare conditions. Performance varied across NHS boards (F1: 86.13-98.13) and improved with patient age (F1: 77.6 for <50 vs. 91.7 for 70+), demonstrating the influence of local reporting conventions and age-specific linguistic variation. Notably, reports for younger patients tend to contain more hedged or uncertain language, which presents a particular challenge for current NLP systems [11] and avenues for future work.

Conclusion

GS-BrainText addresses a significant gap in UK clinical text resources, providing the first multi-site, expertly annotated brain imaging report dataset for UK healthcare contexts. Unlike recent large-scale brain imaging datasets relying on LLM-generated or NLP-system output labels [12,13], it offers gold standard expert annotations suitable for rigorous NLP development and evaluation, enabling investigation of linguistic variation (including diagnostic uncertainty expression), multi-site generalisation challenges and the impact of data characteristics on system performance.

Our EdIE-R baseline demonstrates that well-engineered rule-based approaches can achieve competitive performance on clinical phenotyping tasks while delivering interpretable predictions. It also reveals that even within a single national system, linguistic variation across sites and patient demographics poses challenges to out-of-the-box deployment. Organisations adopting NLP tools should therefore conduct local validation before deployment and monitor performance across demographic and clinical subgroups. Available via Generation Scotland's controlled access process, GS-BrainText serves as a vital resource for developing the robust, generalisable and explainable NLP tools that clinical practice demands. Looking ahead, Generation Scotland's multi-omics and deep phenotyping linkage opens opportunities for multimodal research that extends beyond clinical NLP.

Study context

Ethics & Approvals: This research was conducted under ethical approvals for the Generation Scotland cohort. GS has ethical approval for the SFHS study (reference number 05/S1401/89) and 21CGH study (reference number 06/S1401/27) and both studies are now part of a Research Tissue Bank (reference 20-ES-0021). Ethical approval for the GS:SFHS study was obtained from the Tayside Committee on Medical Research Ethics (on behalf of the National Health Service). Ethical approval for the GS:21CGH study was obtained from the Scotland A Research Ethics Committee. Patients/participants provided their written informed consent to participate in this study.

Stakeholder Involvement: GS-BrainText was created by a multidisciplinary clinical and technical NLP team. The Edinburgh Clinical NLP Group conducted PPIE work, engaging patient representatives on the use of clinical free-text data in AI-driven health research, data sensitivity and sharing as part of the Advanced Care Research Centre.

Data Availability: GS-BrainText will be made available upon publication via controlled access through the Generation Scotland website: https://www.ed.ac.uk/generation-scotland/for-researchers/access.

Acknowledgements: We thank the Generation Scotland participants and the Generation Scotland team. We also thank the annotators (Liam Lee, Michael Walsh, Freya Pellie, Karen Ferguson and William Whiteley) of GS-BrainText.

Funding: Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates (CZD/16/6) and the Scottish Funding Council (HR03006) and is currently supported by the Wellcome Trust (216767/Z/19/Z).

Individual authors (see initials in brackets) were supported by:

- Turing Fellowships (B.A. and C.G.) and a Turing project (B.A. and A.C.) funded by the Alan Turing Institute (EPSRC grant EP/N510129/1)
- Legal & General Group as part of the Advanced Care Research Centre (B.A. and A.C.)
- The AIM-CISC project funded by the National Institute for Health Research (NIHR202639; B.A.)
- An MRC Clinician Scientist Award (G0902303; W.W.)
- A Scottish Senior Clinical Fellowship (CAF/17/01; W.W.)

The funders had no role in the conduct of the study, interpretation or the decision to submit for publication. The views expressed are those of the authors and not necessarily those of the funders.

References

1. Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 6(1):317.
2. Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association, 23(2):304–310.
3. Blair H Smith, Harry Campbell, Douglas Blackwood, John Connell, Mike Connor, Ian J Deary, Anna F Dominiczak, Bridie Fitzpatrick, Ian Ford, Cathy Jackson, et al. 2006.

Generation Scotland: The Scottish Family Health Study; a new resource for researching genes and heritability. BMC medical Genetics, 7(1):74.
4. Blair H Smith, Archie Campbell, Pamela Linksted, Bridie Fitzpatrick, Cathy Jackson, Shona M Kerr, Ian J Deary, Donald J MacIntyre, Harry Campbell, Mark McGilchrist, Lynne J Hocking, Lucy Wisely, Ian Ford, Robert S Lindsay, Robin Morton, Colin N A Palmer, Anna F Dominiczak, David J Porteous, and Andrew D Morris. 2012. Cohort profile: Generation Scotland: Scottish family health study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. International Journal of Epidemiology, 42(3):689–700.
5. Hannah Milbourn, Daniel McCartney, Anne Richmond, Archie Campbell, Robin Flaig, Sarah Robertson, Chloe Fawns-Ritchie, Caroline Hayward, Riccardo E Marioni, Andrew M McIntosh, et al. 2024. Generation Scotland: an update on Scotland's Longitudinal Family Health Study. BMJ Open, 14(6):e084719.
6. Beatrice Alex, Claire Grover, Richard Tobin, Cathie Sudlow, Grant Mair, and William Whiteley. 2019. Text mining brain imaging reports. Journal of Biomedical Semantics, 10(Supplement 1):23.
7. Emily Wheater, Grant Mair, Cathie Sudlow, Beatrice Alex, Claire Grover, and William Whiteley. 2019. A validated natural language processing algorithm for brain imaging phenotypes from radiology reports in UK electronic health records. BMC Medical Informatics and Decision Making, 19(1):184.
8. Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 102–107.
9. Andreas Grivas, Beatrice Alex, Claire Grover, Richard Tobin, and William Whiteley. 2020. Not a cute stroke: analysis of rule-and neural network based information extraction systems for brain radiology reports. In The 11th International Workshop on Health Text Mining and Information Analysis at EMNLP 2020, pages 24–37. Association for Computational Linguistics.
10. Dominic Sykes, Andreas Grivas, Claire Grover, Richard Tobin, Cathie Sudlow, William Whiteley, Andrew McIntosh, Heather Whalley, and Beatrice Alex. 2021. Comparison of rule-based and neural network models for negation detection in radiology reports. Natural Language Engineering, 27(2):203–224.
11. Arlene Casey, Emma Davidson, Claire Grover, Richard Tobin, Andreas Grivas, Huayu Zhang, Patrick Schrempf, Alison Q. O'Neil, Liam Lee, Michael Walsh, Freya Pellie, Karen Ferguson, Vera Cvoro, Honghan Wu, Heather Whalley, Grant Mair, William Whiteley, and Beatrice Alex. 2023. Understanding the performance and reliability of NLP tools: a comparison of four NLP tools predicting stroke phenotypes in radiology reports. Frontiers in Digital Health, 5.
12. Charlotte Maschke, Peter Hadar, Yicheng Zhang, Jian Li, Gauri Ganjoo, Andrew Hoopes, Alessandro Guazzo, Aditya Gupta, Manohar Ghanta, Bruce Nearing, Christine Tsien Silvers, Bharath Gunapati, Robert Thomas, Jennifer A. Kim, Shibani S. Mukerji, Adrian Dalca, Sahar Zafar, Alice D. Lam, Emmanuel Mignot, and M Brandon Westover. 2025. The brain imaging and neurophysiology database: Binding multimodal neural data into a large-scale repository. MedRxiv.
13. Michael PJ Camilleri, Dorian Gouzou, Salim Al-Wasity, Muthu RK Mookiah, Maria Valdes Hernandez, Bea Alex, Sotirios A Tsaftaris, Andrew Brooks, Ruairidh MacLeod, Honghan Wu, et al. 2025. A large dataset of brain imaging linked to health systems data: the curation and access to a whole system national cohort from NHS Scotland. MedRxiv.

Predicting Systematic Review Conclusion Change

Ebrahim Alharbi[1,2], Mark Stevenson[1]

[1]School of Computer Science, University of Sheffield, Sheffield, United Kingdom
[2]Department of Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia

1 Introduction

Systematic reviews (SRs) synthesise the best available evidence on clinical questions to underpin treatment guidelines and health policy [1]. With nearly 80 SRs published per day [2] and 1.5 million PubMed articles indexed annually [3], keeping them current is critical but challenging [4]. A survival analysis found roughly a quarter need updating within two years [5].

A small number of studies have explored predicting when a SR's conclusion is likely to change using machine learning [6, 7]. Bashir et al. [7] used a rule-based approach to extract numerical features (coverage score, search-date gap, trial and participant counts) from Cochrane reviews[1] and their updates; their best approach (a random forest classifier) reached 80.8% accuracy. Coverage score, their most predictive feature, is retrospective: it requires participant counts from the completed update, limiting prospective use.

This study augments structured metadata with biomedical language model representations, and evaluates on a substantially larger dataset (3,326 pairs vs 256). Our results demonstrate that biomedical text embeddings capturing the relationship between a review and its newly identified publications, combined with structured metadata, can effectively distinguish reviews in which conclusions will change from those that will remain unaltered. We frame this as a binary classification task: given an original review and the abstracts of its candidate new publications, predict whether the conclusion of the updated review will differ from the original.

2 Methods and Data

We extracted consecutive version pairs from Cochrane intervention reviews (1995–2021), yielding 3,326 pairs from 2,485 unique reviews across 52 clinical domains. 38% of the pairs had changed conclusions, identified by analysing the structured 'What's New' event codes, supplemented by the text-classification rules of Bashir et al. [7] for ambiguous codes (e.g., UPDATE or AMENDMENT). The pairs were split into training and test sets (80/20) at the review level to ensure no review appeared in both sets. Reference list comparisons between versions identified new publications in the updated version.

[1]Cochrane is an international evidence-synthesis organisation publishing structured systematic reviews. `https://www.cochranelibrary.com/about/about-cochrane-reviews`

Four feature groups were computed from each review–update pair, using only the original review and the abstracts of its candidate new publications as input:
Semantic proximity: BioLinkBERT [8] vectors for the review and each relevant new publication, their element-wise difference, and cosine similarity.
Evidence volume: Relevant new publication counts, participant numbers reported in relevant new publication abstracts, and the time gap between the original review's last search date and the most recent new publication year.
Directional consistency: SciBERT-based natural language inference scores classifying each new publication's relationship with the review as consistent or inconsistent.
Outcome overlap: Lexical and embedding-based similarity between stated review outcomes and new abstract content.

We used abstracts only, as they are consistently available unlike full text. Features are combined using a stacking ensemble (logistic regression, random forest, and SVM base learners; logistic regression meta-learner) with cost-sensitive weighting and 5-fold group cross-validation.

3 Results

We re-implemented the feature extraction of Bashir et al. [7] and trained a random forest with their reported hyperparameters on our dataset, using the same review-level split described in Section 2. Table 1 shows held-out test set (n = 678) performance.

Table 1: Test set performance. NPV = negative predictive value.

Model	Acc	F1	Recall	NPV	AUC-ROC
Baseline: Bashir et al. [7]	59.0	58.5	59.7	71.1	0.64
Our ensemble	**70.4**	**69.5**	**70.4**	**79.5**	**0.76**

Our ensemble approach achieves 70.4% accuracy and 0.76 AUC-ROC, representing substantial improvements over the baseline (59.0% and 0.64 respectively). The 80% NPV means the model is correct roughly four in five times when it predicts a review's conclusions will remain unchanged, which is particularly relevant for editorial triage where confidently identifying stable reviews reduces unnecessary update effort.

Feature ablation showed that biomedical embeddings were the most informative group (-4.1% when removed), while entailment features added nothing, suggesting embeddings already capture contradiction signals.

4 Conclusion

These results show that text and metadata carry enough signal to anticipate conclusion change; in practice, candidate new publications would be identified through routine search and screening, supporting editorial triage. The model was tested entirely on Cochrane reviews; future work includes temporal validation, generalisability testing on non-Cochrane reviews, and large language model integration.

5 Study Context

Review data were accessed under a data sharing agreement with the Cochrane Library. We gratefully acknowledge the Cochrane Library for providing access to this data.

Only published metadata and openly available PubMed abstracts were used; no patient data were involved and ethical approval was not required. No competing interests exist for either author. Since we worked only with published review data, patient and public involvement did not arise. No specific funding was received for this work.

References

[1] Knezevic NN, Manchikanti L, Hirsch JA. Principles of Evidence-Based Medicine. In: Essentials of Interventional Techniques in Managing Chronic Pain. Springer; 2024. p. 101–18.

[2] Hoffmann F, Allers K, Rombey T, et al. Nearly 80 Systematic Reviews Were Published Each Day. J Clin Epidemiol. 2021;138:1–11.

[3] Bittermann A, Kobras L. The Landscape of Biomedical Research. iScience. 2024;27(7):110076.

[4] Cumpston M, Flemyng E. Chapter IV: Updating a Review. In: Higgins JPT, Thomas J, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions version 6.5. Cochrane; 2024.

[5] Shojania KG, Sampson M, Ansari MT, et al. How Quickly do Systematic Reviews Go Out of Date? A Survival Analysis. Ann Intern Med. 2007;147(4):224–33.

[6] Bashir R, Surian D, Dunn AG. The Risk of Conclusion Change in Systematic Review Updates. J Clin Epidemiol. 2019;110:42–49.

[7] Bashir R, Dunn AG, Surian D. A Rule-Based Approach for Automatically Extracting Data from Systematic Reviews to Model Conclusion Change Risk. Res Synth Methods. 2021;12(2):216–25.

[8] Yasunaga M, Leskovec J, Liang P. LinkBERT: Pretraining Language Models with Document Links. In: Proc ACL; 2022. p. 8003–16.

LLM reliability in clinical information extraction from ENT electronic health records

Liam Barrett[1], Nikhil Joshi[1], and Nishchay Mehta[1]

[1]UCL Ear Institute, University College London, London, UK

1 Introduction

Extracting structured clinical information from electronic health records (EHRs) at scale remains a significant challenge [1]. Large language models (LLMs) have shown capability in medical knowledge tasks [2], yet their reliability in extracting clinical information from real-world documentation relative to medical professionals is not well established [3]. This is relevant for ENT, hearing and balance conditions, where deep phenotyping is needed for precision therapeutics [4, 5] but remains under-utilised. Building on prior work using simpler neural networks for ENT information extraction [6], we present a systematic evaluation of seven LLMs against fourteen medical professionals in extracting SNOMED-CT coded clinical information from ENT EHRs, using inter-rater reliability metrics and Bayesian non-inferiority testing.

2 Methods and Data

We conducted a cross-sectional study using 98 ENT EHRs from MTSamples, a publicly available clinical documentation resource that may not capture the full variability of real-world EHRs across healthcare systems. Fourteen qualified doctors with ENT experience independently extracted seven categories of clinical information (socio-demographics, signs, symptoms, diagnoses, treatments, risk factors, and test results). Documents were distributed using a modified Hungarian algorithm to maximise annotator pair diversity, with each document reviewed by at least two independent professionals.

Seven LLMs were evaluated: three proprietary (GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro), three open-source (LLAMA 3.1 405B, 70B, 8B), and one locally deployable (Gemma 3 12B). All used a standardised chain-of-thought prompting strategy with one-shot learning. SNOMED-CT code assignment was enforced for both human and LLM annotators. Inter-rater reliability was assessed using Cohen's Kappa for medic-medic (M-M), medic-LLM (M-L), and LLM-LLM (L-L) pairings. Bayesian hierarchical modelling with Beta-distributed likelihoods and weakly informative priors formally tested non-inferiority of M-L relative to M-M agreement at margins of $\delta = 0.05$, 0.10, and 0.15. Classification performance was evaluated using medical professionals' annotations as ground truth. All code and data are available at `https://github.com/evidENT-AI/LLM-EHR-IRR.git`.

3 Results

Cohen's Kappa was 0.752 (95% CI: 0.710–0.794) for M-M, 0.795 for L-L, and 0.391 (95% CI: 0.362–0.420) for M-L pairs. Bayesian analysis estimated posterior mean agreement of 0.813 (95% CI: 0.755–0.860) for M-M and 0.659 (95% CI: 0.633–0.684) for M-L, a difference of 0.154 (95% CI: 0.091–0.209; Figure 1a). Non-inferiority was rejected at all margins ($P(\text{M-L} \geq \text{M-M} - \delta)$: 0.002, 0.045, 0.430 for $\delta = 0.05, 0.10, 0.15$).

Figure 1: **(a)** Posterior distributions of Cohen's Kappa for M-M (red) and M-L (green) agreement. **(b)** Category-specific Kappa differences (M-L minus M-M) with 95% credible intervals.

Agreement varied by category (Figure 1b), with differences ranging from -0.035 (test results) to -0.099 (diagnoses).

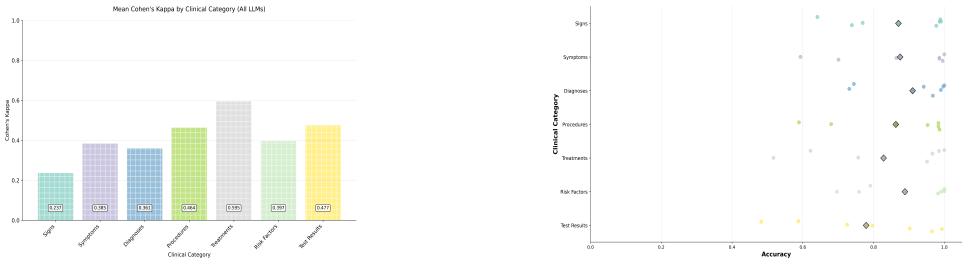


Figure 2: **(a)** Mean M-L Cohen's Kappa by clinical category. **(b)** GPT-4o accuracy by clinical category (diamonds = mean; circles = per-EHR values).

Non-inferiority at $\delta = 0.10$ was achieved only for treatments and test results. Error analysis of GPT-4o across 13,656 annotations showed 97.0% precision, 84.9% recall, and a 7.5% false positive rate, varying by category (Figure 2b). Treatments showed the highest M-L agreement ($\kappa = 0.595$; Figure 2a) and signs the lowest ($\kappa = 0.237$). LLMs extracted more entities per EHR than professionals, particularly for risk factors (7.33 vs 1.01 per EHR).

4 Conclusion

Current LLMs do not achieve inter-rater reliability comparable to medical professionals when extracting SNOMED-coded clinical information from ENT EHRs, as formally demonstrated by Bayesian non-inferiority testing. However, high precision (97.0%) with moderate recall (84.9%) suggests LLMs are well suited for augmentative roles with subsequent human verification. Their tendency to extract more entities than clinicians, particularly for risk factors, indicates potential to complement human expertise by capturing information that clinicians de-prioritise. Two limitations should be noted: MTSamples may not reflect the variability of real-world EHRs across

healthcare systems, limiting external generalisability; and detailed error analysis was restricted to GPT-4o, with future work extending this across all evaluated models to characterise model-specific failure modes. These results provide evidence-based guidance for LLM deployment in clinical documentation workflows.

References

[1] Ashish K Jha, Catherine M DesRoches, Eric G Campbell, Karen Donelan, Sowmya R Rao, Timothy G Ferris, Alexandra Shields, Sara Rosenbaum, and David Blumenthal. Use of electronic health records in us hospitals. *New England Journal of Medicine*, 360(16):1628–1638, 2009.

[2] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[3] Joschka Haltaufderheide and Robert Ranisch. Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, 6(3):e222–e230, 2024.

[4] Nicholas A Lesica, Nishchay Mehta, Federico Engel, Eva Gutierrez-Sigut, Hazel Sheridan, and Anne G M Schilder. Harnessing the power of artificial intelligence to transform hearing healthcare and research. *Nature Machine Intelligence*, 3(10):840–849, 2021.

[5] Lilia Dimitrov, Liam Barrett, Aizaz Chaudhry, Jameel Muzaffar, Watjana Lilaonitkul, and Nishchay Mehta. Uncovering phenotypes in sensorineural hearing loss: A systematic review of unsupervised machine learning approaches. *Ear and Hearing*, 2025.

[6] Nikhil Joshi, Kawsar Noor, Xi Bai, Marina Forbes, Talisa Ross, Liam Barrett, Richard J B Dobson, Anne G M Schilder, Nishchay Mehta, and Watjana Lilaonitkul. Automating the extraction of otology symptoms from clinic letters: a methodological study using natural language processing. *BMC Medical Informatics and Decision Making*, 2025.

Developing a Natural Language Processing Enhanced Liver Transplant Registry

Aditya Borakati[1,3], Jack Wu[1], Christopher Callaghan[2], Miriam Cortes-Cerisuelo[3], James Teo[3], David Wallace[4]
[1] King's College London, London, UK
[2]Guy's and St Thomas' NHS Foundation Trust, London, UK
[3]King's College Hospital NHS Foundation Trust, London, UK
[4]London School of Hygiene and Tropical Medicine, London, UK

Introduction

Liver transplantation is a complex, life-saving procedure for end-stage liver disease. Donor organs are in short supply globally and outcomes following transplant are highly sensitive to donor, recipient and surgical factors. Most countries record data about liver transplants in national registries, to monitor outcomes and for research, to maximise the utility of this constrained, life-saving resource.

These registries typically only record high level structured information such as survival and patient co-morbidities. They fail to capture the wealth of data contained in free-text health records from both recipient and donor, for example, immunosuppressive complications in clinic letters or vascular flow rates in ultrasound scan reports, despite these factors being key drivers of survival [1].

Natural language processing (NLP) has become popularized in healthcare in recent years, however, there has been little adoption in transplantation and limited effort in creating harmonized multicentre datasets from electronic health records (EHRs) with NLP concepts.

In this project, we aim to harmonize liver transplant EHR data, both structured and concepts extracted from unstructured data, into a standardized data warehouse that can be adopted as a blueprint in transplant centres nationally and worldwide. To our knowledge, this would the only such registry worldwide and this novel granular data will allow prediction of outcomes and discovery of new insights on a scale not possible with current registries.

Methods and Data

Data Source

Liver transplant data was sourced from EHR records at King's College Hospital, UK (KCH)- one of highest volume programmes in Europe, from 1988 to present. Donor information was retrieved from the NHS Blood and Transplant registry.

Data was ingested into the CogStack AI data lake instance (based on Elastic, Elasticsearch B.V, Amsterdam, Netherlands) on the KCH network [2].

Data Mapping

A multidisciplinary team of surgeons, hepatologists and intensive care physicians was convened to identify relevant concepts to extract from EHRs including structured concepts and concepts to be extracted from unstructured clinical notes.

These concepts were mapped to Observational Medical Outcomes Partnership (OMOP) Concept IDs, or to new custom Concept IDs where none were found.

Natural Language Processing

A sample of documents for each document type (e.g. imaging report, clinic letter) was annotated with pre-defined concepts of interest by a clinician. These sample were then split training and test sets and used to finetune pretrained models including MedCAT, BERT and GliNER. The best performing models with F1 scores >0.8 and low false positive rates were selected.

Transformation to Common Data Model

The final registry is to be built in the OMOP Common Data Model version 5.4. This is in the form of a recipient-centric relational database [3]. As transplant data is not represented adequately in the base model, we developed a custom Transplant extension table, which records each graft as a row, and donor and graft variables as columns, the recipient ID acts as a foreign key to allow linking to the recipient data. This solution allows use of existing OMOP software for analysis while maintaining ease of analysis for donor and graft concepts.

The search query to identify documents in the data lake, transformation to the final tables and columns, mapping to OMOP Concept IDs, NLP model parameters and metrics were all described in the LinkML data modelling language. LinkML is flexible, to allow addition of new mappings and transformations as EHR data and common data models evolve [4].

Tracking of data files, NLP parameters and metrics was conducted using the Data Version Control (DVC) software, a Git like version control system for tracking data and AI models.

Software

Finding a dearth of appropriate applications for easily mapping EHR variables to the OMOP common data model, we developed our own application with AI assistance.

The LinkML, schemasheets, Zensical and tkinter Python packages were used to develop a software program in Python with a graphical user interface (GUI) to edit variables with a spreadsheet like interface and generate: 1) data dictionary and mapping website 2) LinkML, SQL and other schema definition files

Results

A total of 263 variables were mapped successfully to our custom OMOP data model. These variables included demographic, laboratory, imaging, surgical, medication and co-morbidities.

LinkML Schema Editor Software

The developed GUI application (Figure 1) allows easy entry of concepts, with search of OMOP concepts and automatic entry.

Data Dictionary and Mapping Website

The final website is accessible here: https://a.borakati1.gitlab.io/
A screenshot of an example NLP derived concept is shown below in Figure 2, showing mapping from source, coding and NLP performance.

Conclusion:
We demonstrate a blueprint for transplant units and other clinical specialties to transform their NLP-enhanced EHR data to a structured data warehouse in a standardized common data model.
This will allow description of datasets in a transparent way (in the FAIR principles) and allow linkage of data across centres globally, in addition to linkage with outside datasets such as genomics and national registries such as Hospital Episode Statistics which have previously been transformed to OMOP. This will allow an unparalleled level of granular, multicentre research in transplantation and beyond.

References:

1. Zarrinpar, A. & Busuttil, R. W. Liver transplantation: past, present and future. *Nat. Rev. Gastroenterol. Hepatol.* **10**, 434–440 (2013).
2. Spathis, N., Kraljevic, Z. & Dobson, R. CogStack - integrated information retrieval and extraction architecture. *BMJ Open* **7**, e012569 (2017).
3. Reinecke, I., Zoch, M., Reich, C., Sedlmayr, M. & Bathelt, F. The Usage of OHDSI OMOP - A Scoping Review. *Stud. Health Technol. Inform.* **283**, 95–103 (2021).
4. Moxon, S. A. T. *et al.* LinkML: An Open Data Modeling Framework. *GigaScience* giaf152 (2025) doi:10.1093/gigascience/giaf152.

Study context:

Ethical approval

Health Research Authority ethics (18/LO/2048) and Confidential Advisory Group approvals (23/CAG/0141) have been obtained for collection of patient data without explicit consent.

Funding

Aditya Borakati is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) [Grant reference number EP/Y035216/1] Centre for Doctoral Training in Data-Driven Health (DRIVE-Health) at King's College London, with additional funding from the Roger Williams Institute of Liver Studies at King’s College London and a Fellowship in Clinical Artificial Intelligence from the National Health Service Digital Academy.

Patient and public involvement and engagement

This work was informed by focus groups consisting of 8 liver transplant recipients and their carers from the LISTEN group at King’s College Hospital. All patients were supportive of their anonymized healthcare data being used for research.

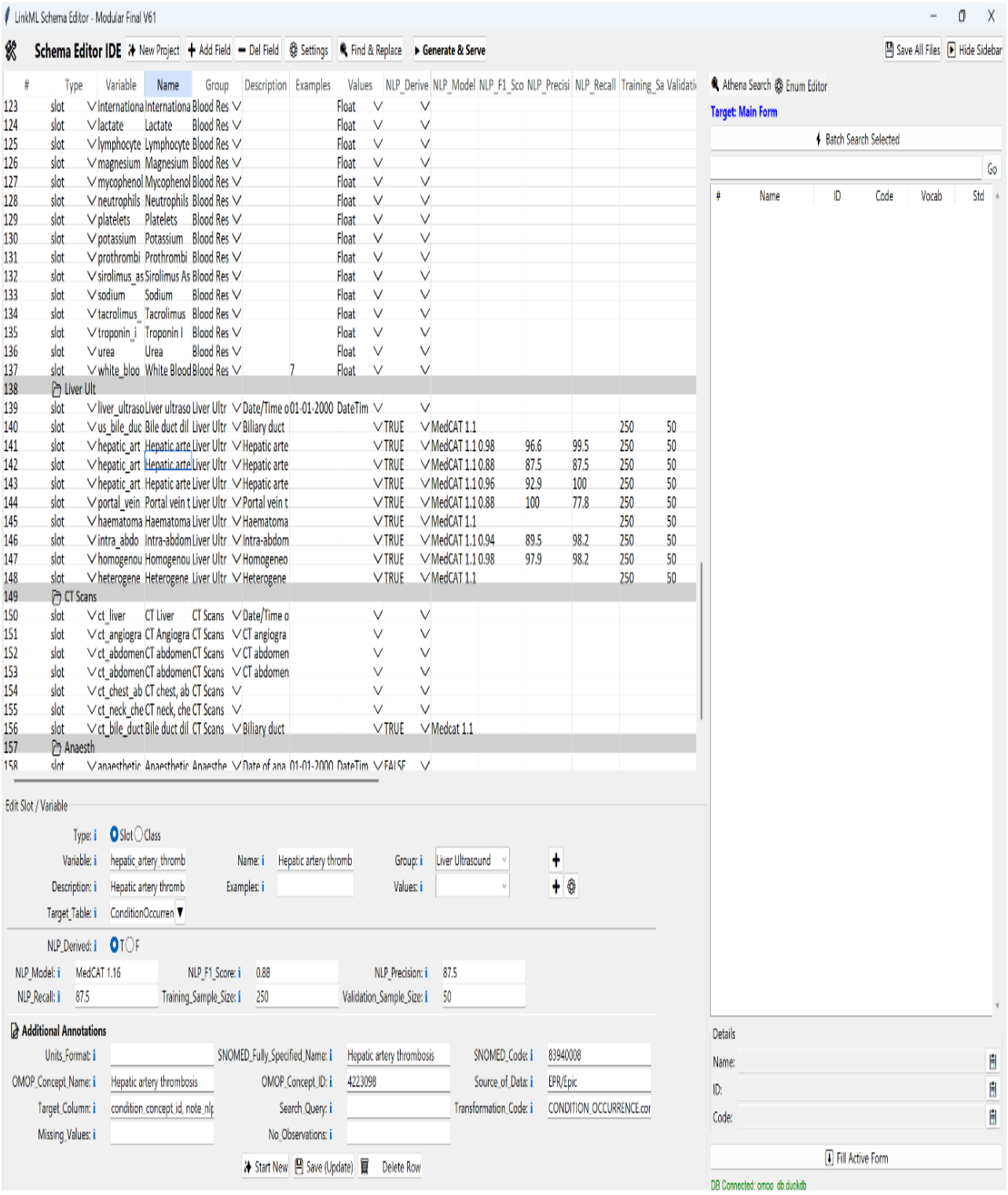


Figure 1- LinkML Schema Editor Application

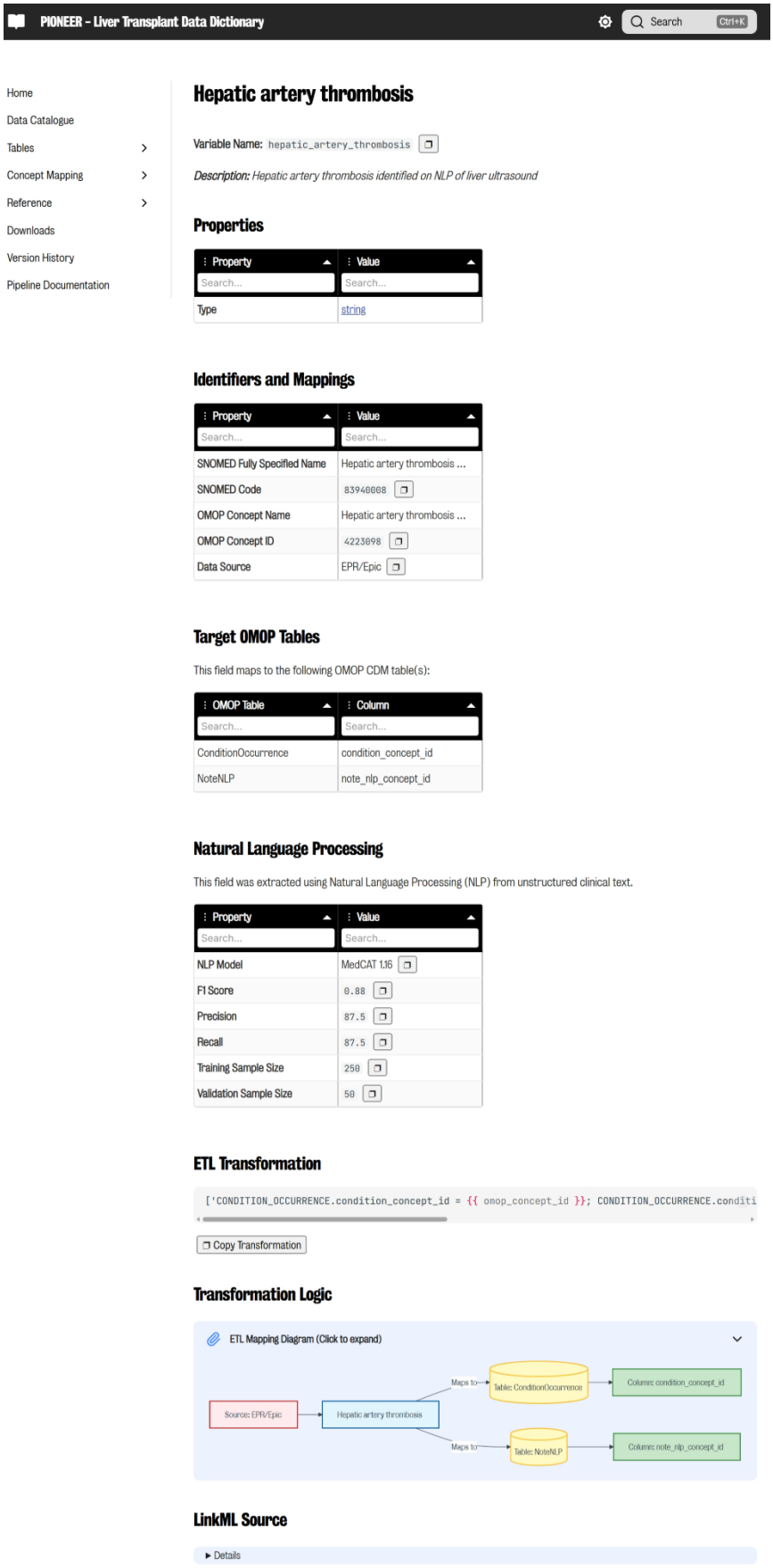


Figure 2- Data dictionary and mapping site, showing example NLP derived variable

Utilization of a fine-tuned BERT model to identify instances of the subject of clinical records experiencing job-loss

David Chandran[1], Alice Broadbent[1], Jyoti Sanyal[2], Robert Stewart[1]
[1] Kings College London, London, United Kingdom
[2] South London and Maudsley Trust, London, United Kingdom

Introduction

The recent growth of Natural Language Processing (NLP) and Transformer-based Deep Learning in the context of clinical records has created new opportunities for extracting and classifying useful information from the free text of clinical records that would otherwise be difficult to acquire and that is not contained elsewhere within the record[1]. One area of particular interest is identifying where the subject of a clinical text had recently experienced a loss of employment. This availability of this information is limited in structured fields of clinical records, making the ability to extract the information from the unstructured (free) text vital for collecting the required information. Within clinical texts, job-loss can be referred to in multiple ways, in various contexts, which can make accurately classifying instances of job-loss a challenging endeavour. This is particularly notable when trying to specifically identify recent instances of job-loss (recent being defined here as within the previous month). The challenge that this research addresses, therefore, is whether a Deep-Learning based approach to clinical text classification could accurately determine the presence of these instances.

Methods and Data

This article presents a novel method for determining instances of job loss within the text of records from within the Maudsley Clinical Record Interactive Search (CRIS) database[2], a large database containing over 500,000 records through use of a BERT model[3]. BERT is a large language model that has been successfully used in classification of data within medical texts[4], including prior CRIS applications[5]. This method involves first using keywords and simple rules to find potential mentions of recent job-less within clinical text, and the surrounding context. The paper then describes using a BERT model, (as well as two other models built using BERT's architecture) to classify these contexts to whether they refer to job-loss or not, to determine whether any can return an acceptable performance.

The first step in creating the model was the development of a dataset that could be used in its training and evaluation. These would be two sets of texts, one that could be used to train the model and another to test the accuracy of the model. The size and variety of data contained within CRIS, made it a very good candidate for the development of the dataset. The first stage in extraction was through a sql query identifying all records that contained a very broad set of keywords that could potentially be related to job loss. From these records the next stage was to identify potential individual instances of job-loss as well as well as the contextual text surrounding them. Another problem in this regard was the broadness of the terms that were used in the extraction. To solve this, simple rules were created to filter out obvious extraneous results through ensuring generic key terms such as "dismissed" were within proximity to terms referring to employment. These rules were implemented using Spacy, an industrial strength NLP library[6], which captured each instance, with contexts of 200 characters around the keyword.

After the dataset was built, a set of annotation rules were created. These instructed the annotators to only positively annotate instances where the job loss was experienced by the subject no later than one month prior to the reference in the instance. Initially a set of 100 instances were randomly selected and annotated by two annotators to determine the level of

inter-annotator agreement. This returned a high IAA value of 89.89%. Following this, the remaining 900 instances in the selection were annotated. This was then split into training data (containing 900 instances) and test data (containing 100 instances). With the completion of the dataset, the training of a model that is able to classify these instances as referring to a recent job-loss or not could begin, using the annotated training data.

The building of the model involved using a pre-trained BERT model and tokenizer (which was required to convert the text data into a numeric form that could be used by a computational model) and then fine-tuning it over the training data with goal of it being able to accurately classify instances that were passed to it, before evaluating it over the test data. Towards this end, three separate base models were looked at. They were the base BERT model[3], RoBERTa a variant of BERT with expanded training and parameters[7], and BioBERT a variant of BERT that was trained specifically over biomedical data[8]. Each of these models was trained over the training data for 30 epochs.

Results

After the models were built, they were evaluated against the annotated test data (see Table 1)

Table 1. Results

Model	Precision	Recall	F1
BERT	87	96	91.5
BioBert	82	96	89
RoBERTa	84	93	88.5

Conclusion

As can be observed from the results, all the BERT models that were utilized were able to classify instances of recent job-loss with a high level of accuracy, with the base BERT model slightly outperforming BioBERT and RoBERTa. This illustrates the efficacy of using a fine-tuned BERT model for this task. Using this method over a rules-based method has allowed for a substantial amount of time and labour. In addition, the BERT method is not computationally expensive and does not require a substantial amount of computational resources to deploy. This gives it a substantial advantage over potential GPT-based LLM approaches, which would have much higher resource and time costs. With the model now having been tested, it can now be used for further research on CRIS data, to determine the prevalence of recent job-loss on either the entirety of CRIS, or over cohorts as the situations require. In terms of future work there are two areas that can be taken forward. First, in terms of improving the job-loss model itself, BERT classification models can be built that find additional features and characteristics related to job-loss, such as instances of historic or repeated job loss. In addition, BERT models can be used to identify other traumatic life events contained in the free text of clinical records that relate to the subject. One area of future work that is being explored in that area is identifying instances of subject bereavement, which would represent an expansion of the scope of the project into a wider exploration of identifying instances of traumatic events that a present within unstructured clinical texts.

Study Context

The Clinical Records Interactive Search (CRIS), which contained the pseudonymised data that was used in the study, was funded by The National Institute for Health and Care Research (NIHR) Maudsley Biomedical Research Centre (BRC) at South London and Maudsley Trust

(SLaM) and King's College London (KCL). Access to the required data was subject to ethical and legal approval by the CRIS Oversight Committee.

References

1. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association. 2018 Oct;25(10):1419-28.

2. Perera G, Broadbent M, Callard F, Chang CK, Downs J, Dutta R, Fernandes A, Hayes RD, Henderson M, Jackson R, Jewell A. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) case register: current status and recent enhancement of an electronic mental health record-derived data resource. BMJ open. 2016 Mar 1;6(3):e008721.

3. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. InProceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) 2019 Jun (pp. 4171-4186).

4. Li F, Jin Y, Liu W, Rawat BP, Cai P, Yu H. Fine-tuning bidirectional encoder representations from transformers (BERT)–based models on large-scale electronic health record notes: an empirical study. JMIR medical informatics. 2019 Sep 12;7(3):e14830.

5. Chaturvedi J, Velupillai S, Stewart R, Roberts A. Identifying mentions of pain in mental health records text: a natural language processing approach. arXiv preprint arXiv:2304.01240. 2023 Apr 3.

6. Honnibal M, Montani I, Van Landeghem S, Boyd A. spaCy: Industrial-strength natural language processing in Python.

7. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234-40.

8. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Multimorbidity Patterns in Adults with Intellectual Disability and Digital Tools for Exploring Care Experiences

Georgina Cosma[1*], G. Thomas Jun[1], Mikel W. Lekuona[1], and Saul Albert[1]

[1]Loughborough University, UK
Corresponding author: g.cosma@lboro.ac.uk

1 Introduction

Multiple long-term conditions (MLTCs), defined as the presence of two or more chronic physical or mental conditions, represent a major challenge for healthcare systems [1, 2]. Adults with intellectual disability experience earlier onset and higher prevalence of chronic illness compared with the general population, contributing to significant health inequalities and reduced life expectancy [3, 4]. Previous studies have reported high prevalence of conditions such as epilepsy, mental illness, and respiratory disease in people with intellectual disability. However, less is known about how these conditions cluster and develop over time, particularly across age groups and between sexes. Understanding patterns of multimorbidity may support improved screening, preventive care, and clinical decision-making. This study examines the prevalence and temporal associations of multiple long-term conditions in adults with intellectual disability in England using linked primary and secondary care data. In addition, we present digital tools developed in this work to explore multimorbidity patterns and complementary care experience evidence, including simulated and reported care experiences generated through a conversational AI system.

2 Methods and Data

This observational study used CPRD GOLD (July 2023 build) primary care records linked to Hospital Episode Statistics (HES) admitted patient care and outpatient data, together with Office for National Statistics (ONS) mortality data. Individuals with intellectual disability were identified between 01/01/2000 and 30/06/2023. Participants were adults aged 18 years and over with at least one recorded diagnosis of intellectual disability and at least one year of up-to-standard registration. Patients were included if they had two or more long-term conditions. The full available longitudinal record for each patient within the study period was used to identify diagnoses and examine temporal relationships between conditions. The final cohort included 13,051 adults with intellectual disability (7,123 males and 5,928 females). Forty long-term conditions were identified using Read codes in primary care data and ICD-10 codes in hospital records. Patients were stratified by sex and by age at first diagnosis into three groups: under 45 years, 45–64 years, and 65 years and over. Statistical associations between condition pairs were assessed using Fisher's exact test with odds ratios and 95% confidence intervals, and false discovery rate correction was applied to account for multiple comparisons. We developed an interactive tool for exploring patterns of multiple long-term conditions in adults with intellectual disability using aggregated linked healthcare data. The tool analyses co-occurrence and temporal relationships between conditions and generates visual networks illustrating how diseases cluster across the life course (see Fig. 2). These analyses support exploration of potential disease progression pathways and investigation of multimorbidity patterns.

3 Results

Most diagnoses occurred before age 45 for both sexes (73.4% in females and 73.7% in males). The mean number of conditions increased with age, indicating accumulation of multimorbidity across the life course. Among females, the mean number of conditions increased from 2.58 before age 45 to 3.99 among those aged 65 years and over, while in males the increase was from 2.31 to 3.76. Mental illness was the most prevalent condition among females (41.21%), while epilepsy was the most prevalent condition among males (36.30%). Hypertension affected approximately 28% of both sexes, while reflux disorders and chronic airway diseases were also common. Age-specific patterns were observed.

In younger adults (aged <45 years), epilepsy and mental illness were the most frequent diagnoses. In the 45–64 age group, cardiometabolic conditions including hypertension and diabetes became more prevalent. Among adults aged 65 years and over, multimorbidity patterns were dominated by chronic kidney disease, dementia, cardiac arrhythmias, and heart failure. Strong associations between conditions were observed. One of the strongest and most frequent associations occurred between diabetes and hypertension. Mental illness also showed repeated associations with insomnia, chronic pain conditions, and reflux disorders and often preceded other conditions within multimorbidity trajectories.

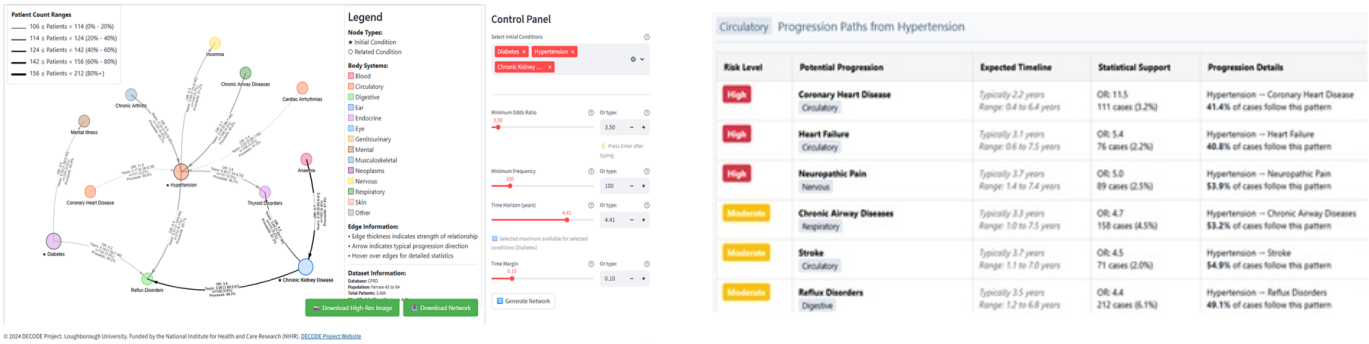


Figure 1: Interfaces of the multimorbidity analysis tool.

4 Conversational AI Care Experience Tool

We developed a conversational AI tool designed to support the exploration and analysis of healthcare interaction scenarios. The system can generate simulated conversations between clinicians and patients using predefined personas that represent different clinical roles, patient characteristics, and care contexts. In addition to fully simulated interactions, the platform supports AI–human conversations in which either the clinician or patient role can be performed by a human participant while the other role is generated by the AI system. The tool also provides functionality for thematic analysis of conversation transcripts in order to identify recurring concerns, communication challenges, and care experience themes. To support safety evaluation, we integrated a structured safety analysis framework that enables systematic assessment of conversational content. Evaluation components were implemented to assess how well the AI system follows instructions and responds appropriately when interacting about potentially sensitive healthcare issues.

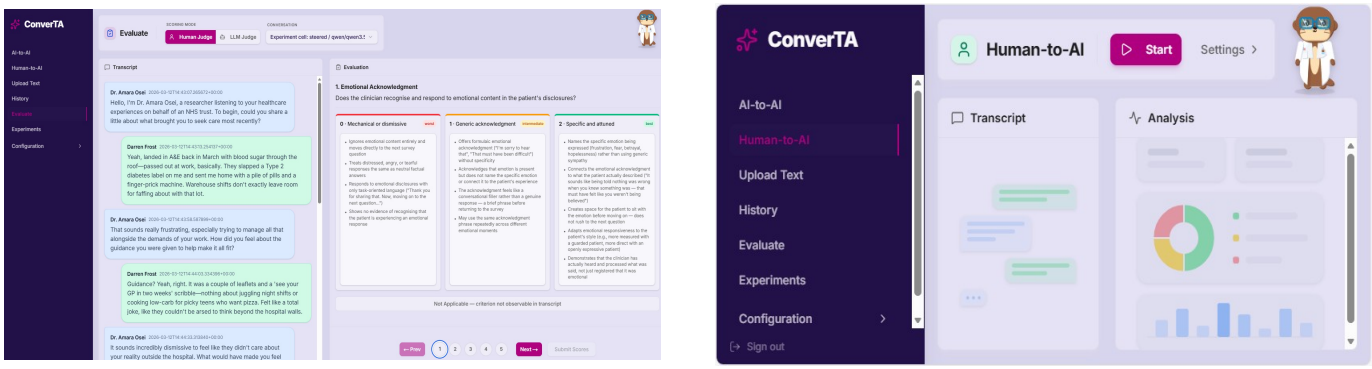


Figure 2: Interfaces of the conversation generation and analysis tool.

5 Conclusion

Adults with intellectual disability experience a substantial burden of multiple long-term conditions, with clear age-specific and sex-specific patterns. Neurological and mental health conditions are more prominent earlier in adulthood, while cardiometabolic and age-related conditions become more common later in life. Understanding these patterns may support earlier detection and more targeted screening strategies. To complement this population-level analysis, we developed digital tools to support exploration of multimorbidity patterns and care experiences. These include an interactive tool for analysing relationships between long-term conditions using aggregated statistical data and a conversational AI system designed to collect and analyse care experiences and generate simulated healthcare conversations involving individuals with intellectual disability and their carers. Together, these approaches aim to provide a broader perspective on healthcare risks, care barriers, and lived experiences affecting this population.

6 Study context

Acknowledgments. Data-driven machinE-learning aided stratification and management of multiple long-term COnditions in adults with intellectual disabilitiEs (DECODE) project (NIHR203981) is funded by the NIHR AI for Multiple Long-term Conditions (AIM) Programme. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. This work uses data provided by patients and collected by the NHS as part of their care and support. We also want to acknowledge all data providers who make anonymised data available for research.

Role of the funding source. The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Ethics. This study has been approved by the CPRD Independent Scientific Advisory Committee (protocol number: 22_001840).

Data Availability. The individual-level patient data used in this study cannot be shared publicly and are not available for redistribution by the authors, as access was granted solely for the purposes of this study under the terms of the CPRD data access agreement (protocol number: 22_001840). Researchers wishing to conduct similar studies may apply independently for access through CPRD (`https://www.cprd.com`) subject to their own approvals. The Read code lists, condition definitions, chronicity criteria, and epidemiological analysis plan are publicly available via the Open Science Framework at `https://doi.org/10.17605/OSF.IO/KT5FY`.

References

[1] Dambha-Miller H, Cheema S, Saunders N, Simpson G. Multiple Long-Term Conditions (MLTC) and the Environment: A Scoping Review. International Journal of Environmental Research and Public Health. 2022;19(18):11492.

[2] Valabhji J, Barron E, Pratt A, Hafezparast N, Dunbar-Rees R, Turner EB, et al. Prevalence of multiple long-term conditions (multimorbidity) in England: a whole population study of over 60 million people. Journal of the Royal Society of Medicine. 2024;117(3):104-17.

[3] Cooper SA, Hughes-McCormack L, Greenlaw N, McConnachie A, Allan L, Baltzer M, et al. Management and prevalence of long-term conditions in primary health care for adults with intellectual disabilities compared with the general population: A population-based cohort study. Journal of Applied Research in Intellectual Disabilities. 2018;31(Suppl 1):68-81.

[4] Kinnear D, Morrison J, Allan L, Henderson A, Smiley E, Cooper SA. Prevalence of physical conditions and multimorbidity in a cohort of adults with intellectual disabilities with and without Down syndrome: cross-sectional study. BMJ Open. 2018;8(2):e018292.

Agentic system for research specific real world data quality checks

Joseph Cronin[1], Keiran Tait[1], Robert Dürichen[1]
[1] Arcturis Data Ltd, Kidlington, Oxfordshire, UK

Introduction

Real world data (RWD), including electronic health records, is playing an increasingly central role in clinical research and the generation of real-world evidence (RWE). However, RWD is primarily collected to support direct patient care rather than research, resulting in data that may be incomplete, inconsistent, or poorly aligned with the needs of a specific research study. As a result, robust data quality assessment is essential to ensure that datasets are fit for purpose.
A range of established tools exist for generic data quality assessment, many of which provide valuable generic checks and descriptive summaries such as OHDSI's Achilles [1] and Achilles Heel or the PEDSnet Network Data Quality (NDQ) toolkit [2]. However, when it comes to project-specific checks, researchers typically resort to bespoke analyses—a time-consuming and resource-intensive process.
The emergence of large language models (LLMs) and autonomous agents introduces new possibilities for automating research-specific data quality controls. For example, Li *et al.* [3] investigated LLMs for anomaly detection in tabular data and Li *et al.* [4] introduced a multi-agent framework for automatic bias detection.
We present the first insights into ArcVAL (Arcturis VALidation), an agentic system designed to perform research-specific data quality checks and report results on 2 of the 3 main components, information extraction and data analysis. ArcVAL analyses RWE study protocols, extracts key requirements, maps them to standardised clinical concepts, and automatically evaluates the presence, plausibility, and completeness of the corresponding data elements. By aligning data quality assessment directly with study objectives, ArcVAL aims to improve the efficiency, transparency, and reliability of RWE studies.

Methods and Data

ArcVAL is implemented as an agentic system composed of three main components: Information Extraction (IE), Translator (T), and Data Analysis (DA), see Figure 1. Each component operates as an independent single- or multiagent workflow with a distinct objective. This modular design improves accuracy through greater control of individual components and enhances flexibility by enabling the use of task-specific LLMs (e.g., lightweight models for in-SDE data analysis). The IE component analyses study-agent workflow with a distinct objective. The IE component analyses study specific documents, such as protocols or statistical analysis plans (SAPs), to identify data elements relevant to the research questions, for example diagnosis codes or demographic criteria. This component is implemented as a sequential agentic workflow, with different agents responsible for extracting elements from specific data categories. The Translator component maps the extracted data elements to one or more target ontologies used within the dataset of interest. This standardisation step ensures that study requirements are aligned with the underlying data representation, enabling consistent downstream analysis. Finally, the DA component performs research-specific data quality assessments. It dynamically generates SQL queries tailored to the standardised concepts and predefined quality tests and executes them against the dataset. The DA component follows a planner–executor–verifier (PEV) approach, enabling iterative refinement of the generated SQL code based on query results (up to a maximum of five iterations). The data quality tests are defined in natural language to ensure flexibility across different database schemas, with a database schema description provided to the planning agent as contextual input.

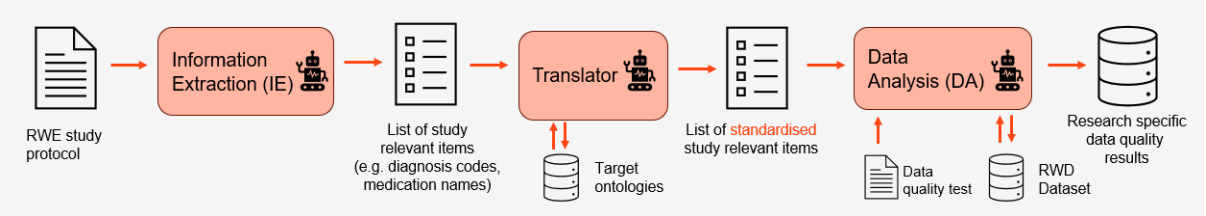


Figure 1. High-level system diagram of ArcVAL with its main components

The IE component was developed using three internally available statistical analysis protocols (SAPs), with an average length of 92 pages and additional tabular appendices. The objective was to extract all data elements belonging to the following categories: demographics, diagnoses, procedures, medications, laboratory tests, and clinical observations. The IE component uses Claude 3.7 Sonnet for the individual agents. Ground truth annotations were generated manually.

The DA component was evaluated on an anonymised prostate cancer dataset from the Arcturis RWD Network Research Database, to which two NHS trusts contributed (13,449 patients, time range: 2006-2025). Performance was assessed based on the ability of the system to generate correct SQL code and the number of PEV cycles required across 16 predefined data quality tests. These tests were targeting diagnosis codes and were divided into two groups of equal size. Simple tests involved single-table aggregations (e.g. counting patients with a given diagnosis code or estimating first or last occurrence of a code), whereas complex tests required multi-table joins and advanced logic including temporal constraints and demographic stratification (e.g. age or gender stratification of a defined patient cohort). The LLaMA-3-8B model was included as a representative example of a smaller, open-source LLM, with the aim of assessing whether tasks could be performed in more constrained environments, such as secure data environments (SDEs) of the NHS.

Results

The evaluation of the IE component on the 3 SAPs reveals an average F1-scores (& standard deviation) for extraction of demographic information, diagnosis codes, procedure codes, clinical observations, laboratory tests and medication names: 0.6 (0.12), 0.97 (0.02), 0.86 (0.25), 0.63 (0.11), 0.97 (0.03), 0.97 (0.02) respectively. The higher F1-scores observed for diagnoses, procedures, medications, and laboratory tests are largely attributable to the fact that these data elements are typically represented using well-defined and structured coding systems, such as ICD-10 codes for diagnoses and OPCS-4 codes for procedures. In addition, the study protocols often specified a large number of such codes, which frequently appeared in structured tabular formats. In contrast, demographic variables and some clinical observations achieved lower F1-scores, reflecting both the smaller number of items typically specified and their greater semantic variability. For example, a single concept such as age may be expressed in multiple forms (e.g. age, date of birth, or year of birth), increasing ambiguity during extraction. Moreover, certain variables present edge cases—such as body mass index (BMI)—where categorisation is inherently ambiguous, leading to challenges in consistently assigning them to demographics versus clinical observations.
The evaluation of the DA component shows that the larger Claude model can generate accurate SQL code for all simple data quality tests (100% accuracy), both with and without the use of SQL templates, in the first PEV iteration. For complex quality tests, Claude achieves an accuracy of 37.5% without templates and 87.5% when templates are provided. In contrast, the LLaMA model can’t generate correct SQL code for any quality tests without templates; when templates are provided, performance increases, achieving accuracies of 50% and 25% for simple and complex tests, respectively.

Conclusion

We present the first insights into ArcVAL, a novel agentic system designed to perform research specific- data quality checks for RWE studies. The evaluation of the IE component demonstrates that key study relevant- data elements can be extracted from study documents with high accuracy. The IE component is currently optimised to a specific document type (SAPs) used in this study. Further optimisation is required to make it robust to different documents formats.
The evaluation of the DA component shows promising results when using a big LLMs, especially for complex data quality checks when supported by SQL templates Future work will focus on scalable template generation (e.g. via RAG-based template libraries), incorporation of human-in-the-loop feedback, and fine-tuning of smaller models to improve performance on complex queries in constrained environments.

Study context

An anonymised dataset for this study was provided by the Arcturis Real-World Data Network Research Database under REC approval 24/YH/0164. Our thanks to all participating NHS trusts who provide data to the database, and to the patients and members of the public who advise and support Arcturis.

References

[1] DeFalco F, Ryan P, Schuemie M, Huser V, Knoll C, Londhe A, et al. Achilles: Achilles Data Source Characterization. R package version 1.7.2. 2023. https://github.com/OHDSI/Achilles (accessed February 19, 2026).

[2] Razzaghi H, Dickinson K, Wieand K, Bailey C. Network Data Quality (NDQ) n.d. https://pedsnet.github.io/ndq/ (accessed February 19, 2026).

[3] Li A, Zhao Y, Qiu C, Kloft M, Smyth P, Rudolph M, et al. Anomaly Detection of Tabular Data Using LLMs. ArXiv 2024.

[4] Li H, Ma MD, Huang J, Weng Z, Wang W, Zhao J. BIASINSPECTOR: Detecting Bias in Structured Data through LLM Agents. ArXiv 2025.

Predicting Clinical Outcomes for Patients with Mental Illness using NLP on Electronic Health Records

Jakob G. Damgaard Msc[1,2,3], Kenneth Enevoldsen Msc PhD [3], Sara Kolding Msc[1,2,3], Frida Hæstrup Msc[1,2,3], Erik Perfalk MD PhD[1,2,3], Andreas A. Danielsen MD PhD[1,2], Søren D. Østergaard MD PhD[1,2]

[1]Department of Affective Disorders, Aarhus University Hospital, Denmark

[2]Department of Clinical Medicine, Aarhus University, Denmark

[3]Aarhus NLP group, Center for Computing Humanities, Aarhus University, Denmark

Introduction

The digitalisation of electronic health records (EHRs)—coupled with the advancements in modelling and data representation methods—has created a promising environment for the integration of predictive artificial intelligence into healthcare as decision support tools that may help clinicians issue more personalised and timely treatment. We have previously been successful in predicting a range of clinically relevant outcomes such as mechanical restraint (1, 2), involuntary admission (3), need for electroconvulsive treatment (4), diagnostic progression to schizophrenia or bipolar disorder (5), and initiation of clozapine treatment (unpublished) at the level of the individual patient, using only routine clinical data from EHRs from psychiatric services.

In psychiatry specifically, details and nuances regarding patient states, symptoms and treatments are often more accurately captured in free-text clinical notes than in conventional structured variables. Our preliminary studies are built around conventional machine learning classification frameworks trained on multi-modal feature vectors derived from the EHRs. For these models, clinical notes were tabularized using simple statistical methods like TF-IDF or a pre-trained sentence transformer. Despite these simple free-text representations, models trained exclusively on note-derived features performed comparably to models trained on expert-curated features from structured EHR data. Given these encouraging results, we now aim to extend this line of work and take advantage of our unique catalogue of different prediction models by systematically exploring the further potential for enhancing clinical prediction modelling using natural language processing (NLP).

Specifically, we will compare the predictive performance of NLP models spanning an increasing spectrum of complexity and exhaustively attempt to map out benefits and pitfalls of different methods for different prediction tasks. This study aims to deepen our understanding of the value and feasibility of utilizing free-text clinical notes for clinical models, ultimately contributing to their integration in clinical practice.

Methods and Data

The study builds data from the PSYchiatric Clinical Outcome Prediction cohort encompassing routine clinical EHR data from all individuals with at least one contact to the Psychiatric Services of the Central Denmark Region (6). The dataset covers >120.000 adult patients and contains comprehensive data on all patient contacts, including clinical notes, diagnoses, medications, lab values and coercive measures.

For the study, we aim to implement and explore the following approaches to modelling clinical notes for prediction tasks:

- *Classifiers (e.g., XGBoost) trained on feature vectors based on the term frequency of predefined psychopathology-related words*
- *Classifiers trained on TF-IDF-derived feature vectors*
- *Classifiers trained on embeddings from a state-of-the-art (SOTA) off-the-shelf embedding model*
- *Classifiers trained on embeddings from a SOTA embedding model fine-tuned on the clinical outcome prediction tasks*
- *Prompt-based classification using SOTA large language models, including both general-purpose and healthcare-specific models*

These methods will be trained on the prediction tasks outlined in the introduction, and their performance will be compared with that of the multi-modal models presented in the original studies. Evaluation will focus on predictive performance, temporal and geographical stability, and interpretability. In particular, we will assess whether the models maintain robust performance across different time periods and hospital sites and evaluate how the degree of interpretability and complexity of the factors underlying model predictions may affect their clinical utility and applicability.

Results and conclusion

The study is still in its early phases and results are not yet available. Preliminary results are expected by medio 2026.

Study context

The development of AI in the field of healthcare is sensitive and should be undertaken with utmost precaution. Any algorithms developed in this project are intended solely as support tools for clinicians and never as a means for automating decisions regarding patients. Furthermore, analyses will be conducted to uncover and mitigate any potential biases in the algorithms.

The study is funded by Independent Research Fund Denmark, The Lundbeck Foundation and the Danish Agency for Digital Government. The study was approved by the Legal Office of the Central Denmark Region in accordance with the Danish Health Care Act §46, Section 2 (1-45-70-60-25). The Danish Committee Act exempts studies based only on EHR data from ethical review

board assessment. Handling and storage of data complied with the European Union General Data Protection Regulation. The project is registered on the list of research projects having the Central Denmark Region as data steward.

According to Danish law, the patient-level data for this study cannot be shared. The code for all analyses will be available at:

https://github.com/Aarhus-Psychiatry-Research/psycop-common/tree/main

There is no public nor patient involvement in this study.

Conflicts of interest:

A. A. D. has received a speaker honorarium from Otsuka Pharmaceutical. S. D. Ø. received the 2020 Lundbeck Foundation Young Investigator Prize and S. D. Ø. owns/has owned units of mutual funds with stock tickers DKIGI, IAIMWC, SPIC25KL, DKIEUIXBNP and WEKAFKI, and owns/ has owned units of exchange traded funds with stock tickers BATE, TRET, QDV5, QDVH, QDVE, SADM, IQQH, IQQJ, USPY, EXH2, 2B76, IS4S, OM3X, MCHI and EUNL. The remaining authors declare no competing interests.

References

1. Danielsen, A. A., Fenger, M. H., Østergaard, S. D., Nielbo, K. L., & Mors, O. (2019). Predicting mechanical restraint of psychiatric inpatients by applying machine learning on electronic health data. Acta Psychiatrica Scandinavica, 140(2), 147-157.

2. Kolding, S., Damgaard, J. G., Bernstorff, M., Hansen, L., Østergaard, S. D., & Danielsen, A. A. (2025). Development and Evaluation of Machine Learning Models to Predict Mechanical Restraint and Related Coercive Measures in Hospital Psychiatry. medRxiv, 2025-12.

3. Perfalk, E., Damgaard, J. G., Bernstorff, M., Hansen, L., Danielsen, A. A., & Østergaard, S. D. (2024). Predicting involuntary admission following inpatient psychiatric treatment using machine learning trained on electronic health record data. Psychological Medicine, 54(15), 4348-4361.

4. Hansen, L., Damgaard, J. G., Lundin, R. M., Danielsen, A. A., & Østergaard, S. D. (2025). Predicting the need for electroconvulsive therapy via machine learning trained on electronic health record data. medRxiv, 2025-06.

5. Hansen, L., Bernstorff, M., Enevoldsen, K., Kolding, S., Damgaard, J. G., Perfalk, E., ... & Østergaard, S. D. (2024). Predicting diagnostic progression to schizophrenia or bipolar disorder via machine learning applied to electronic health record data. medRxiv, 2024-07.

6. Hansen L, Enevoldsen KC, Bernstorff M, Nielbo KL, Danielsen AA, Østergaard SD. The PSYchiatric clinical outcome prediction (PSYCOP) cohort: leveraging the potential of electronic health records in the treatment of mental disorders. Acta Neuropsychiatr. 2021;1–8. https://doi.org/10.1017/neu.2021.22

Creating A Gold-Standard Annotated Epilepsy EHR Dataset

Joe Davies[1], Beata Fonferko-Shadrach[2], Ben Holgate[1], Arron Lacey[2], Huw Strafford[2], Owen Pickrell[2], and Mark P. Richardson[1]

[1]King's College London, London, England
[2]Swansea University, Swansea, Wales

1 Introduction

Epilepsy is a common neurological disorder affecting over 51 million people worldwide [1]. It can be managed with medication and lifestyle factors, usually facilitated by specialist clinics that generate electronic health records (EHRs), which largely consist of free-text clinic letters. As EHRs are a vital resource for clinical and research use, extracting information from them is a promising avenue for research.

Natural Language Processing (NLP) can help analyse EHRs by parsing human speech. Recent NLP applications include: extracting seizure frequency [2], current medications [3], and long-term seizure patterns [4] from clinic letters.

Most NLP methods leverage Large Language Models (LLMs), which encode information about term usage and sentence relational information. Some medical-purpose LLMs include MedGemma [5], Meditron [6], and BioMistral [7]. While useful, these models typically require tuning which necessitates finding gold-standard, expert-annotated data. Available medical data corpora include MIMIC-IV [8], NPA-CP [9], and EHRCon [10], but there's a lack of freely available clinic visit EHRs due to identifiable patient data constraints. This work aims to create a corpus of annotated, anonymized EHRs with easy-to-follow guidelines. This corpus could then be used to fine-tune LLMs by training the LLM on the EHRs with labels from the annotations.

2 Methods and Data

We use 3000 EHRs: 1500 from King's College London Hospital (KCLH) and 1500 from Guy's and St Thomas' Trust (GSTT). The different centres have similar methods for writing clinic notes, reducing the need for data harmonisation across sites. Data was anonymised using Anon-Cat [11]. This is a transformer-based model that removes identifiable information such as: home/work/email addresses, names, dates of birth, telephone numbers etc. Anonymisation was validated before annotation was completed by visual inspection of the EHRs, as well as by those annotating. No issues were found.

Data was annotated by 7 users: 4 clinicians and 3 data scientists with epilepsy domain experience. This was done through the annotation platform called Markup [12]. Markup provides

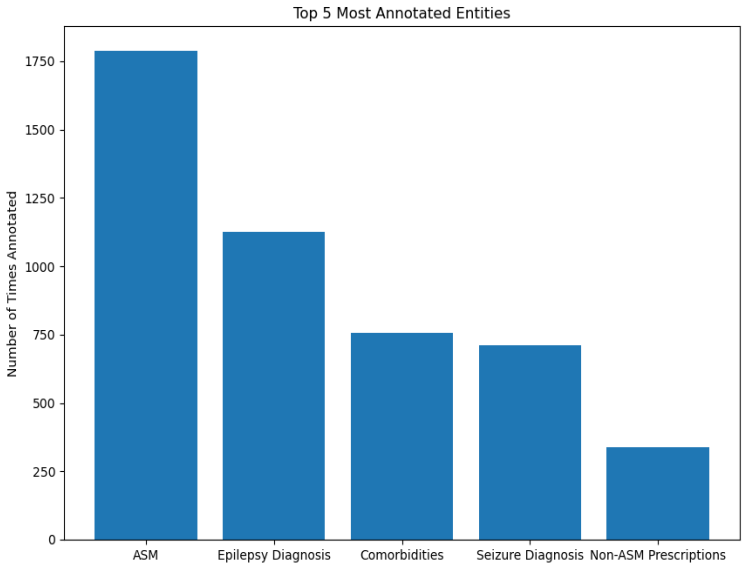


Figure 1: The top 5 most annotated entities in 1350 clinic letters. Counts represent the total unique annotations over all annotations in a given group of entities. Here, ASM is anti-seizure medication.

annotators with a list of entities to annotate, each with their associated attributes. The attributes are either selected from a drop-down menu of options, or inputted by the user. These entities and attributes are set by a configuration file which is manually loaded onto the system, allowing for bespoke guidelines to be used. This configuration file is of .json format, and was developed using Markup's GUI tool for this purpose. After annotations are completed, the data can be exported in the form of a .ann file (using the Brat Standoff format [13]), which can be read as an ordinary text file.

3 Results

So far, 1350 total documents have been annotated, with 750 of these being double annotated. The most annotated entities are those concerned with current anti-seizure medications (ASMs) being annotated by at least one annotator a total of 1789 times. The least annotated are rare or incorrect annotations such as extremely premature births, or mistakes in annotation input. The top 5 most annotated can be found in figure 1. Inter-annotator agreement of double-annotated EHRs is undertaken using Krippendorff's Alpha [14], calculated by:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1}$$

where: D_o and D_e are observed and expected disagreement respectively. Krippendorff's Alpha is chosen because we use both binary (present or not-present) and non-binary (drug dose, for

Letters	Mean Alpha Score
1-150	0.66
301-450	0.74
451-600	0.75
601-750	0.74
1351-1500	0.71

Table 1: Krippendorff Alpha scores for different spans of letters in 150 letter chunks. Work in progress.

example) annotations.
First, the raw annotations are grouped. This is to reduce sparsity, minimise artificial disagreement between closely related labels, and evaluate agreement at the level of clinically meaningful concepts rather than highly granular annotations. Next, we calculate α for these categories over each letter annotated, getting an overview of the agreement over all the selected letters. Then a mean value is found. The results can be found in table 1. So far we have a mean of 0.72. This shows a promising level of agreement at this early stage. The most agreed upon annotations are ASMs and their associated dosages, often reaching a score of 1.00. The least agreed upon is information related to dates given, such as the onset of epilepsy or the year that seizure freedom was achieved. These are often negative, indicating this information is difficult to agree upon and is, perhaps, more ambiguous. Work is ongoing.

4 Conclusion

In summary, we use commonly annotated entities to provide a gold-standard dataset of annotated clinical letters. A total of 1350 letters are annotated by a team of clinicians and data scientists (750 of which are annotated twice). The top 5 most numerous categories of entities annotated include information pertaining to current ASMs, diagnoses related to epilepsy and comorbidities, and any other prescriptions a patient may be on. This work is ongoing and will be updated as more annotations are added and analysis performed.

5 Study context

De-identification was performed using in-house, verified de-identification software, AnonCAT. Data will not be made available as per the conditions of a data sharing agreement. The configuration file, however, may be given on request. There are no conflicts of interest at this time. Work is funded with a generous grant from the Medical Research Council.

References

[1] GBD Epilepsy Collaborators. Global, regional, and national burden of epilepsy, 1990–2021: a systematic analysis for the global burden of disease study 2021. *The Lancet Public Health*, 2025.

[2] Ben Holgate et al. Fine-tuning llms to extract epilepsy seizure frequency data from health records. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, 2025.

[3] Shichao Fang et al. Extracting epilepsy-related information from unstructured clinic letters using large language models. *Epilepsia*, 2025.

[4] Kevin Xie et al. Long-term epilepsy outcome dynamics revealed by natural language processing of clinic notes. *Epilepsia*, 2023.

[5] Khaled Saab et al. Capabilities of gemini models in medicine, 2023.

[6] Zeming Chen et al. Meditron-70b: Scaling medical pretraining for large language models, 2023.

[7] Yanis Labrak et al. Biomistral: A collection of open-source pretrained large language models for medical domains, 2024.

[8] Alistair et al Johnson. MIMIC-IV. *PhysioNet*, October 2024. Version 3.1.

[9] Ziming Wei et al. Predictors of Hospital Onset Infection: A Matched Retrospective Cohort Dataset. *PhysioNet*, November 2025. Version 1.0.0.

[10] Yeonsu Kwon et al. EHRCon: Dataset for Checking Consistency between Unstructured Notes and Structured Tables in Electronic Health Records. *PhysioNet*, March 2025. Version 1.0.1.

[11] Zeljko Kraljevic et al. Validating transformers for redaction of text from electronic health records in real-world healthcare. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 544–549, 2023.

[12] Samuel Dobbie et al. Markup: A web-based annotation tool powered by active learning. *Frontiers in Digital Health*, 2021.

[13] Pomtus Stenetorp et al. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, 2012.

[14] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Inc., 2019.

Synthetic free-text healthcare data: informing research design through public involvement

Rebecca Goulding[1], Gail Davidge[1], Sarah Markham[2], Warren Del-Pinto[1], Brain McMillan[1], Goran Nenadic[1]
[1] University of Manchester
[2] Public contributor

Introduction

Real-world free-text healthcare data is inherently sensitive, with risks to patient privacy. Several efforts have been invested in automated pseudo-anonymisation of free-text data [1,2], but residual re-identification risks [3] have often resulted in significant restrictions in accessing such data for research, even if data is placed within Trusted Research Environments (TREs). Synthetic free-text data are emerging [4] as an alternative and trustworthy solution, where data is generated through an AI process that ensures privacy and task-specific data relevance. While public views on synthetic structured data have been explored[1], little is known about perspectives specific to synthetic free-text data.

As part of the FORTRESS-TeHR: Federated, Open and Reliable TREs for Synthetic Textual Healthcare Records project, we are piloting an approach to generate, validate and document realistic synthetic free-text healthcare data that can be securely accessed within TREs for training and validation of AI models [5]. The project will test whether synthetic free-text healthcare data can meaningfully support research while reducing privacy risks. It will develop validation frameworks to help TRE operators, regulators, and researchers understand when synthetic healthcare free-text data is appropriate and where its limitations lie. The project is co-designed with technical specialists, clinicians and patients to ensure trust in synthetic free-text healthcare data and its use in AI training. Here, we reflect on how public involvement has informed and shaped the research and development of the FORTRESS-TeHR project.

Methods and Data

Public involvement is embedded throughout the project. We have established a diverse Public Advisory Panel (PAP), which meets at key decision points to shape the requirements and expectations around privacy, quality, acceptable use, and public benefit. The PAP consists of 12 members (6 female, 5 male, 1 non-binary) geographically spread across England. Half of the public contributors are from non-white British background. Four contributors are between 18 and 35 years, five between 36 and 56 years, and three above that age. The PAP includes 5 carers to ensure diverse perspectives throughout the project lifecycle.

Working with our Public contributor co-investigator, we have organised three meetings with the PAP to discuss specific questions, share views and help shape the research and research outputs.

Meeting 1 - setting the scene. We discussed the idea of generation of synthetic free-text healthcare data to train AI models and how to balance privacy and usefulness of such data, including any benefits and risks to patients and the public.

Meeting 2 – generating the data. We discussed how synthetic free-text healthcare data can be generated, what the key steps are and how such data may be validated. We also discussed the potential formats of the research outputs from the project: for example, the options to share a model(s) that generate such data, or only the data itself.

[1] E.g. https://www.cardiff.ac.uk/centre-for-trials-research/research/studies-and-trials/view/Discussing-Data

Meeting 3 – using the data. We will discuss the deployment and access to synthetic free-text healthcare data, including acceptable ways share a model and/or dataset, how datasets should be described (meta-data) so that they are FAIR (Findable, Accessible, Interoperable, Reusable), and how the process can be managed, transparent and trustworthy in the eyes of the public.

We have already conducted meetings 1 (online) and 2 (face-to-face in Manchester), while meeting 3 (online) is planned for end of March 2026.

Results

The discussions in the PAP meetings are informing the design of technical solutions, the data quality assessment framework and data deployment and access policy. These discussions are documented to ensure transparency on how public involvement influenced the project design and execution.

For example, when discussing the overall idea of using real-world free-text data to guide the development of synthetic free-text data, our public contributors highlighted a requirement to clearly define the purpose of the synthetic free-text data, and what expected use cases are. This has led us to ensure that the used real-world data is relevant for the purpose, represents the right population (e.g. local UK data or even regional) and has temporal relevance (e.g. old real-world data should not be used for training models that need contemporary datasets). There was a clear suggestion to investigate whether synthetic free-text healthcare data generation could have an opt-in rather than opt-out approach.

Some of our PAP contributors noted that the terminology used around synthetic data in general (e.g. sandpits) can indicate that the data is for “playing around with” rather than “serious” research or development. As a consequence, we have adjusted our terminology and will be making recommendations to the wider research community.

Discussions of the option to access either models or synthetic datasets challenged out initial plans. We are now considering sharing both with different users via different licensing options.

Discussions of the nature and seriousness of errors that may appear in synthetic free-text healthcare data have helped us prioritise which validation approaches need to be implemented as minimum, and which ones are more important from the clinical fidelity perspective. Given that patients are increasingly offered to access to their healthcare data, a suggestion was to include patient assessment in the validation framework. Debating the quality of data, the PAP contributors suggested that there is a need to be clear about what the value of synthetic data is and that the role and value of real-world data for specific research tasks (including testing and validation) need to be evident.

Conclusion

Our public involvement activities have been an extremely valuable activity that have helped us clarify public expectations and inform the design requirements. Discussions with our PAP have shaped and adjusted some of the early decisions as part the project lifecycle, including how to specify (and check) the real-world input data, what and who needs to be involved in the validation, and how to provide access to research outputs. Our experience demonstrates the benefits of working with members of the public, alongside other stakeholders, to guide project design and deployment, no matter how technical they may seem.

Study context

This work has been designed as a public involvement activity to shape the development, evaluation and deployment of synthetic free-text healthcare data. The work has been funded by DARE UK through the FORTRESS-TeHR project. Public contributors have been compensated for their time and direct expenses.

References

1. Kovačević A, Bašaragin B, Milošević N, Nenadic G (2024). De-identification of clinical free text using natural language processing: A systematic review of current approaches, Artificial Intelligence in Medicine 151, https://doi.org/10.1016/j.artmed.2024.102845.
2. Falis M, Gruber F, McInerney S, Casey A. (2025). Evaluating LLMs' Potential to Identify Rare Patient Identifiers in Patient Health Records. Studies in health technology and informatics. 327, p. 874-875
3. Ford E, Pillinger S, Stewart R, Jones K, Roberts A, Casey A, Goddard K, Nenadic G. (2025). What is the patient re-identification risk from using de-identified clinical free text data for health research? AI and Ethics (2025).
4. Falis M, Gema AP, Dong H, Daines L, Basetti S, Holder M, Penfold RS, Birch A, Alex B (2024). Can GPT-3.5 generate and code discharge summaries? Journal of the American Medical Informatics Association 31 (10) 2024, https://doi.org/10.1093/jamia/ocae132
5. Wu Y, Schlegel V, Del-Pinto W, Nandakumar S, Zahid I, Sun Y, Omar U, Jasmine A, Kaliya-Perumal AK, Tham C, Connors G, Bharath A, Nenadic G (2025). Term2Note: Synthesising Differentially Private Clinical Notes from Medical Terms (to appear). 10.48550/arXiv.2509.10882.

Extraction of Antidepressant Response from Primary-Care Free-Text Data with Large Language Models

Matúš Falis[1,2], Alice Eaton[2], Michael Holder[2], Kieran Sweeney[2], Matthew H. Iveson[1], Samuel McInerney[2], Franz Gruber[2], Emily L. Ball[1], Heather C. Whalley[1], Arlene Casey[2]

[1] Institute for Neuroscience and Cardiovascular Research, University of Edinburgh, Edinburgh, UK
[2] Centre for Population Health Sciences, Usher Institute, University of Edinburgh, Edinburgh, UK

Introduction

In this study we aim to extract antidepressant phenotypes from primary-care notes using a Large Language Model (LLM) - Llama-3.1-70B [1] in a few-shot-prompting setup. Extraction of such concepts in psychiatry using natural language processing (NLP) focuses mostly on secondary-care data, likely due to the availability of datasets and platforms, such as CRIS [2]. However, patients treated for depression in secondary care constitute less than a quarter of all patients treated for depression in the UK [3]. Much of the treatment process, including prescribing antidepressants, is handled within primary care. Thus, primary-care free text is an important but under-researched area for understanding prescribing patterns, and treatment response in depression.

In depression treatment, the physician's impressions of the patient are primarily recorded in free text [4]. Vaci et al. [4] with the aid of clinical experts proposed a labelling schema for medication response, and related information of interest, in depression and applied it to secondary-care data. They achieved high F1 scores for some entity types (e.g., medication name (0.9) and dosage (0.93)), while reporting poor performance on others, notably ones corresponding to outcomes – medication response and adverse drug reaction (≤0.3). Thus, there is potential for improvement in predicting outcomes and applying it to primary care where most of the treatment happens. Taking an existing labelling schema [2], we adapted it to an LLM setup and applied it to primary-care notes from the Lothian region of Scotland, hosted within DataLoch, to extract antidepressant phenotype information.

Here we present preliminary work including our translation of a previous approach [4] into LLMs, our initial impressions of the benefits of LLMs in this setting, and our adjustments to the labelling schema [4]. We further share our plans for creating a gold standard; model evaluation; future use of LLM outputs for training smaller models; and integrating the extracted phenotypes in the wider scope of the AMBER project [5].

Methods and Data

We have employed Llama3.1-70B [1] to extract entities corresponding to entity types presented in previous work [4] from free text of a sample primary-care encounters in Lothian coded with ReadCodes relevant to depression. Extracted entities included patient history, medication, adverse reaction, symptoms, treatment response and

clinical questionnaire results – each further split into subtypes (e.g., medication included medication name, dosage, frequency, etc.). A prompt was designed for each entity type, with some further broken down into multiple prompts based on theme (e.g., symptoms split into physical, mood, and suicidality). Each prompt tasked the LLM to determine the presence of entities, provide a relevant sentence from the input as evidence to support the prediction, and, finally, the literal string of the entity. The task was extended to present the entities grouped into relations of pre-determined format (e.g., medication-response; symptom-diagnosis). Well-formed outputs of the LLM were sufficient for identifying the entity's indices within the input text. Some modifications were made to the original labelling schema, with certain entity-specific attributes – e.g., severity, or the person to whom an entity relates – changed into entities of their own. Of the originally presented attributes we retained temporality and negation. We further added an entity type for treatment adherence. The LLM produced formatted textual output which we post-processed into entities and relations.

Results

While labelled data enabling evaluation and quantitative analysis is yet to be produced, we can comment on the behaviour observed thus far in the output of the LLM. In some cases (often with extremely short inputs) the model hallucinated, providing outputs and evidence absent from the input. Further to this, the model sometimes classified each sentence in the input as an instance of each entity type relevant to the prompt. When valid predictions were made, they were usually correct, even for entity types that allow for variability, such as patient history (including relevant references to temporality – e.g., differentiating between a divorce happening in the past year versus three years ago as history relevant to current episode versus past history adulthood history). Mentions of response and medication (including changes in prescribing, e.g., dose increase) were identified correctly, which is encouraging for the aims of the project and AMBER [5] as a whole. For the diagnosis entity type, the model tended to predict diagnoses based on mentions of symptoms, rather than focusing exclusively on the exact mentions of psychiatric diagnoses in the input document. The model generally benefitted from the prompt conditioning the entity recognition task on a document-level theme prediction (binary classification on whether the topic of a given entity type is discussed within the document, followed by individual entities if true). Some of the issues described could be addressed with post-processing.

Conclusion

We have presented an ongoing project focusing on extraction of antidepressant response from primary care data. While the gold-standard labels for evaluating model performance are yet to be produced, initial outputs show promising behaviour on entity types of major interest, especially medication-related data and medication response. Some of the undesirable model behaviour can be addressed with post-processing, while other issues may require further prompt engineering. Once the model has been evaluated, we intend to employ it on a larger amount of data to produce silver-standard

labels which will be used in training smaller architectures (e.g., BERT [6]) in more specialised tasks, such as medication extraction, and response extraction.

Study context

This study has been conducted as part of the AMBER (Antidepressant Medications: Biology, Exposure & Response) [5] project funded by a Wellcome Trust Mental Health Award. MHI is supported by the Wellcome Trust (220857/Z/20/Z; 226770/Z/22/Z, 104036/Z/14/Z; 216767/Z/19/Z) and by a Research Data Scotland Accelerator Award (RAS-24-2). ELB also acknowledges additional support from MQ; Transforming Mental Health [MPSIP\30].

This work uses data collected by the NHS as part of their provision of patient care and support. Following exceptional approval by NHS Lothian, the analysis was completed on an extract of patient data in a Trusted Research Environment managed by the DataLoch service: a partnership of the University of Edinburgh and NHS Lothian. One goal of the overall programme of work is to develop a scalable process for DataLoch in which clinical free-text can be de-identified for potential use within approved research projects. The data used for our development project cannot be requested in its raw form.

References

1. Grattafiori, Aaron, et al. "The llama 3 herd of models." arXiv preprint arXiv:2407.21783 (2024).
2. Stewart, Robert, et al. "The South London and Maudsley NHS foundation trust biomedical research centre (SLAM BRC) case register: development and descriptive data." BMC psychiatry 9.1 (2009): 51.
3. Strawbridge, Rebecca, et al. "Care pathways for people with major depressive disorder: A European Brain Council Value of Treatment study." European Psychiatry 65.1 (2022): e36.
4. Vaci, Nemanja, et al. "Natural language processing for structuring clinical text data on depression using UK-CRIS." Evidence Based Mental Health 23.1 (2020).
5. AMBER: Antidepressant Medications: Biology, Exposure & Response [Internet]. [place unknown: publisher unknown]. [date unknown]. Available from:https://www.kcl.ac.uk/research/amber-antidepressant-medications-biology-exposure-response
6. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019.

Systematic Review of Natural Language Processing for Extracting Psychiatric Medication Response from Clinical Free Text (Antidepressants, Antipsychotics, and Mood Stabilisers)

Matúš Falis[1,2], Matthew H. Iveson[1], Samuel McInerney[2], Franz Gruber[2], Emily L. Ball[1], Heather C. Whalley[1], Arlene Casey[2]

[1] Institute for Neuroscience and Cardiovascular Research, University of Edinburgh, Edinburgh, UK
[2] Centre for Population Health Sciences, Usher Institute, University of Edinburgh, Edinburgh, UK

Introduction

In this study we aim to analyse research in retrieving outcomes from medication (response and adverse reaction) in the context of free-text notes in psychiatry (with a focus on antidepressants and antipsychotics). Depression is a major global cause of disease burden. Treatment-resistant depression is a phenotype constituting a failure to respond to at least two different adequately delivered antidepressants. Precision psychiatry requires a better understanding of treatment response, yet response is difficult to measure using structured clinical data. Most depression cases are handled in primary care (including prescribing antidepressants) with the clinician's impressions being primarily recorded in free text. To motivate further research in this area, we have thus conducted a systematic review of the use of Natural Language Processing (NLP) methods for retrieving response to psychiatric medications from clinical free text.

Methods and Data

We searched the literature focusing on keywords relating to the conjunction of NLP, psychiatric medication (antidepressants, antipsychotics and mood stabilisers), Electronic Health Records, and response. The initial search was conducted in January 2025 and was further updated via rerunning the search and enhancement via citation snowballing in June 2025. We identified 1081 publications for screening with 15 publications reaching the extraction phase. We focused on extracting features of the datasets (source in the clinical pathway; availability; language of the free text), the definitions and labelling of response (treatment response/adverse reaction; drug families considered; representation of response – e.g., binary/ternary; etc.), machine learning/NLP methods used, and evaluation methods. We employed the PROBAST tool [1] to evaluate the risk of bias with each publication having been evaluated by two researchers. We removed PROBAST questions irrelevant to NLP, as in previous work [2]. When applying PROBAST we adjusted the PROBAST criteria based on whether the paper focused primarily on technical NLP or epidemiology.

Results

Of the 15 extracted publications only one involved two datasets, with the rest working with a singular dataset. Only one publication focused exclusively on primary-care data, with secondary care being dominant (9 exclusively secondary-care, 3 combined with a different clinical source). This is despite psychiatric conditions (especially relating to

mood) being commonly treated in primary care and likely due to the landscape of available datasets (9 publications included instances of datasets accessible with approvals, with only 1 instance of freely available data due to being synthetic). There was a relatively even split between publications focusing on medication response and adverse reaction (Fig. 1); with four publications using a different outcome variable (e.g., symptom improvement – a proxy to medication response). Similar to some spaces within the scope of Clinical NLP (e.g., cardiology [5] and adverse drug event detection [6]), rule-based methods tend to dominate the landscape (Fig. 2). Classical (pre-deep-learning) NLP appears sparsely, with one instance in 2011 and three since 2020. Deep-learning techniques appeared in three instances, all since 2020 constituting a relatively late adoption even for Clinical NLP standards (e.g., compared adoption in ICD coding in 2018 [7]). LLMs have seen adoption since 2021, including BERT [8] architectures, and the more recent and powerful GPT-3.5.

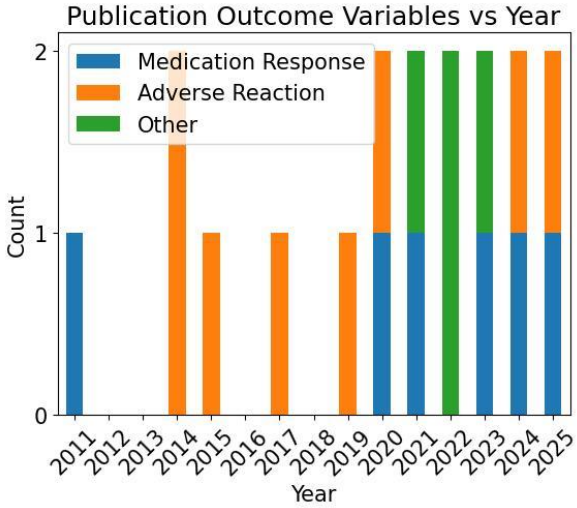


Fig.1: Prevalence of outcome variables in the extracted publications

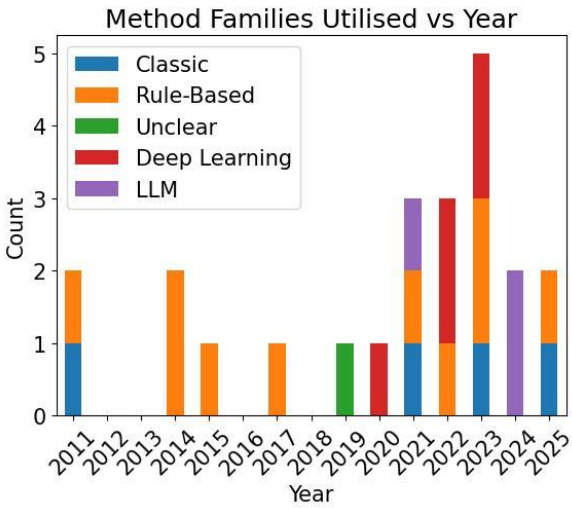


Fig.2: Prevalence of methods in the extracted publications

Conclusion

We found that most datasets are either available with approvals or unavailable to most researchers. This hinders desirable procedures, such as study replication, or external validation. Most studies came from secondary-care sources, which are not representative of the most common setting for prescribing antidepressants and antipsychotics – primary care. As recording between primary and secondary care is different (specialty of the clinician, size and detail of the notes), results acquired on secondary-care data may not translate into primary-care settings. There is a split in focus in these studies with some leaning more into epidemiology while others focusing more on the NLP task leading to different levels of engagement with NLP methods.

Study Context

This study has been conducted as part of the AMBER (Antidepressant Medications: Biology, Exposure & Response) [3] project funded by a Wellcome Trust Mental Health Award. MHI is supported by the Wellcome Trust (220857/Z/20/Z; 226770/Z/22/Z, 104036/Z/14/Z;

216767/Z/19/Z) and by a Research Data Scotland Accelerator Award (RAS-24-2). ELB also acknowledges additional support from MQ; Transforming Mental Health [MPSIP\30]. No patient data was used in this study.

References

1. Wolff, Robert F., et al. "PROBAST: a tool to assess the risk of bias and applicability of prediction model studies." Annals of internal medicine 170.1 (2019): 51-58.

2. Guellil, Imane, et al. "Natural language processing for detecting adverse drug events: A systematic review protocol." NIHR Open Research 3 (2024): 67.

3. AMBER: Antidepressant Medications: Biology, Exposure & Response [Internet]. [place unknown: publisher unknown]. [date unknown]. Available from:https://www.kcl.ac.uk/research/amber-antidepressant-medications-biology-exposure-response

4. Strawbridge, Rebecca, et al. "Care pathways for people with major depressive disorder: A European Brain Council Value of Treatment study." European Psychiatry 65.1 (2022): e36.

5. Turchioe, Meghan Reading, et al. "Systematic review of current natural language processing methods and applications in cardiology." Heart 108.12 (2022): 909-916.

6. Golder, Su, et al. "Leveraging natural language processing and machine learning methods for adverse drug event detection in electronic health/medical records: a scoping review." Drug safety 48.4 (2025): 321-337.

7. Mullenbach, James, et al. "Explainable prediction of medical codes from clinical text." Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers). 2018.

8. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019.

Appendix A: Search Example

Example query for *SCOPUS: TITLE-ABS-KEY (((natural AND language AND processing) OR (natural AND language AND processing AND systems) OR (text AND mining) OR (information AND extraction)) AND ((antipsychotic) OR (mood AND stabiliser) OR (mood AND stabilisers) OR (mood AND stabilizer) OR (mood AND stabilizers) OR (neuroleptic)) OR ((antidepress*) OR (depress*)) AND ((ehr) OR (emr) OR (electronic AND health AND record) OR (electronic AND medical AND record) OR (clinical AND "notes")))*

To what extent can existing toxicity- and sentiment-oriented language models reliably distinguish between psychologically harmful and constructive negative comments?

Omotayo Faluyi
Loughborough University, United Kingdom

Introduction

Social media platforms play a key role in contemporary communication but also expose users to large volumes of negative commentary which has been linked to increased stress, anxiety, and depressive symptoms, particularly among the young and impressionable populations [1]. To mitigate these risks, platforms increasingly rely on automated moderation systems to detect toxic or abusive language. However most existing approaches frame toxicity as a largely binary or multi-label problem focused mainly on identifying profanity, hate speech or explicit harassment.

Prior work in abusive language detection has primarily focused on categorizing harmful content such as hate speech, offensive language, or personal attacks (e.g.,[4,10,11]). While related research has explored dimensions such as constructiveness and helpfulness in online discourse [5,7,9], these are typically treated as separate tasks rather than explicitly contrasted with harmful language.As a result, the current moderation systems often conflate the different forms of negativity, failing to distinguish between comments that are critical but constructive and those that are psychologically harmful.

This distinction is important. Constructive criticism–negative feedback designed to support improvement–can promote learning and engagement[2], whereas psychologically harmful language is intended to demean, shame or emotionally harm the recipient. Systems that fail to distinguish between these forms risk suppressing legitimate discourse or inadequately protecting users.

Recent advances in large language models (LLMs) have improved contextual understanding in toxicity detection. However, it remains unclear whether models trained on surface-level cues can reliably distinguish constructive from harmful negativity, as this depends on intent and psychological impact.

This paper reframes moderation as a **three-class classification problem**: supportive/positive, constructive negative, and psychologically harmful negative.

Contributions

1. It introduces a well-being-oriented reframing of toxicity detection that explicitly separates constructive and psychologically harmful negativity.

2. It provides a comparative empirical evaluation of classical and transformer-based models on this task.
3. It demonstrates a systematic limitation of current models, namely their tendency to confuse harmful comments with constructive criticism.

Methods and Data

We define a three-class classification task:

- **Supportive/positive (S):** non-toxic, affirming comments
- **Constructive (C):** critical but improvement-oriented feedback
- **Harmful (H):** language intended to insult or harm

Evaluation focuses on distinguishing **constructive vs harmful**, the core challenge.

Dataset

We use the Civil Comments dataset[4], a large-scale corpus of user-generated online comments annotated for toxicity attributes. This dataset is particularly suitable because it contains naturally occurring discussions with a wide spectrum of toxicity, including subtle and context-dependent language.

Original toxicity labels are mapped to the three proposed categories based on prior work in constructiveness and abusive language detection [5,6. This mapping enables the study of nuanced distinctions within negative language rather than treating toxicity as binary.

Preprocessing

Text preprocessing includes:

- Lowercasing
- Removal of URLs and punctuation
- Whitespace normalization.

Classical models use TF-IDF representations with unigrams and bigrams.Tranformer models are trained on raw tokenized text, truncated or padded to 128 tokens.

Models

Classical: Logistic Regression, Linear SVM, Random Forest, XGBoost
Transformers: DistilBERT-base-uncased, RoBERTa-base

Transformers are fine-tuned for two epochs using AdamW (learning rate 2×10^{-5}).

Evaluation Metrics

Accuracy, macro F1-score, per-class precision/recall, and confusion matrices.
Macro F1 ensures balanced evaluation.

Results

Model	Accuracy	Macro F1	Constructive F1	Harmful F1
Logistic Regression	72%	0.68	0.70	0.65
Linear SVM	74%	0.70	0.72	0.68
Random Forest	70%	0.66	0.68	0.63
XGBoost	73%	0.69	0.71	0.67
DistillBERT	80%	0.75	0.77	0.73
RoBERTa	81%	0.77	0.79	0.75

Transformer models outperform classical baselines, with RoBERTa achieving the best performance. Compared to DistilBERT, RoBERTa improves accuracy by 1% and macro F1 by 0.02, with the largest gain in harmful comment recall (+3%).

Confusion Matrix

DistilBERT

Predicted	Constructive	Harmful
Constructive	6106	857
Harmful	1178	1859

RoBERTa

Predicted	Constructive	Harmful
Constructive	6216	747
Harmful	1147	1890

A consistent pattern emerges across both models: **harmful comments are more frequently misclassified as constructive than vice versa**. DistilBERT misclassifies 1,178 harmful comments as constructive, compared to 857 constructive comments as harmful. Similarly,

RoBERTa misclassifies 1,147 harmful comments as constructive, compared to 747 constructive comments as harmful.

This corresponds to approximately **38–39% of harmful comments being misclassified**, indicating that both models are more likely to **underestimate harm**. These findings suggest that transformer models struggle to capture psychological intent, particularly when harmful language resembles constructive feedback.

Discussion

Although transformer models improve performance, they remain limited in capturing **pragmatic intent**. They rely on surface-level linguistic cues, which are insufficient for distinguishing constructive criticism from harmful language when both share similar wording.

This limitation highlights a broader challenge in content moderation: psychological harm is not solely a function of explicit toxicity, but also of intent, tone and context. As a result, systems trained on traditional toxicity labels may fail to provide reliable moderation in nuanced cases.

Conclusion

This study demonstrates that existing toxicity- and sentiment-oriented language models can distinguish between constructive and psychologically harmful negative comments only to a limited extent.

These findings underscore the need for moderation systems that go beyond surface-level toxicity detection and explicitly model **communicative intent and psychological impact**.

Future work will explore:

- Instruction-tuned and intent-aware models
- Incorporation of contextual and conversational features
- The effectiveness of large language models (LLMs) and advanced AI systems on this task, particularly their ability to reason about intent, tone, and context beyond surface-level linguistic features.

This work supports the development of moderation systems that reduce harm while preserving constructive discourse.

Study Context

This study uses publicly available, anonymised data from the Civil Comments dataset. No new data was collected, and no personal identifying information was processed. Ethical approval was not required. No external funding was received, and there are no conflicts of interest.

References

1. Smith A, Brown R, Johnson L. Online hostility and youth mental health: a longitudinal study. *Br J Dev Psychol*. 2025;43(2):201–219.
2. Cambridge Dictionary. Constructive criticism [Internet]. Cambridge University Press; 2026 [cited 2026 Jan 10]. Available from: https://dictionary.cambridge.org
3. Duchene C, Jamet H, Guillaume P, Dehak R. A benchmark for toxic comment classification on the Civil Comments dataset. *Proc Papers with Code*. 2023.
4. Wulczyn E, Thain N, Dixon L. Ex machina: personal attacks seen at scale. In: *Proceedings of the 26th International World Wide Web Conference (WWW)*; 2017. p. 1391–1399.
5. Park J, Lim S, Choi Y. Classifying constructive comments. *First Monday*. 2016;21(6).
6. Nguyen LT, Nguyen KV, Nguyen NLT. Constructive and toxic speech detection for social media comments. *arXiv preprint*. 2021;arXiv:2106.12345.
7. Kim SM, Pantel P, Chklovski T, Pennacchiotti M. Automatically assessing review helpfulness. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2006. p. 423–430.
8. Ghose A, Ipeirotis P. Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics. *IEEE Trans Knowl Data Eng*. 2011;23(10):1498–1512.
9. Pitler E, Nenkova A. Finding thoughtful comments from social media. In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*; 2008. p. 699–707.
10. Pavlopoulos J, Malakasiotis P, Androutsopoulos I. Quantifying the impact of context on the quality of manual hate speech annotation. *Nat Lang Eng*. 2023;29(2):437–468.
11. Poletto F, Basile V, Sanguinetti M, Bosco C, Patti V. Problems of varying annotations to identify abusive language in social media content. *Nat Lang Eng*. 2021;27(6):739–757.
12. Cheng S. Dataset augmentation for counteracting bias in toxic comment classification. *Highlights Sci Eng Technol*. 2024;85:1108–1114.
13. Darcy N, Roy S. Violence online: a social media toxicity review. *Violence Gend*. 2017;4(4):210–216.
14. The unappreciated role of intent in algorithmic moderation of abusive content on social media. *Harvard Kennedy School Misinformation Review*. n.d.
15. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: state-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*; 2020. p. 38–45.

Challenges in Generating Realistic Synthetic Clinical Records Using Large Language Models for Evaluating AI Summarizing Tools

Elizabeth Ford[1] Rob Dickinson[1] Edmund Broadhead[2]

[1] Brighton and Sussex Medical School, Brighton, UK
[2] Health Narrator Ltd, UK

Introduction

Healthcare narratives such as referral letters, clinical reports, diagnostic summaries, and discharge documentation contain a large proportion of the contextual information used in clinical decision-making [1]. However, a lengthy set of letters in the patient record can be challenging for a clinician to mentally summarise in the few minutes prior to a patient consultation. These free-text records are therefore an important target for Natural Language Processing (NLP) systems designed to assist clinicians in summarising patient histories, identifying relevant signals, and supporting clinical workflow [2]. However, evaluating and validating such systems requires access to large volumes of patient records.

For technologies intended to be deployed as clinical decision-support tools or medical devices, regulatory frameworks typically require evidence that the system performs reliably across diverse patient populations and clinical contexts. In primary care settings, this requirement presents a particular challenge because general practitioners encounter an extremely wide spectrum of conditions, comorbidities, and diagnostic pathways. Demonstrating robustness therefore requires testing systems against varied and realistic clinical records.

Access to real patient records for such testing is often constrained by privacy regulations, data governance requirements, and institutional approval processes. Synthetic clinical data therefore represents a potential alternative for developing and evaluating healthcare NLP systems [3]. Large Language Models (LLMs) have recently been proposed as tools capable of generating synthetic clinical narratives that resemble real medical documentation [4].

This study explores the feasibility of using LLMs to generate a broad corpus of synthetic patient letters and reports, representing a broad range of general practice patients. Specifically, the work investigates whether LLM-generated clinical narratives can reproduce the structural, stylistic, and informational characteristics of real-world medical documentation sufficiently to support realistic evaluation of summarisation systems.

The 3 objectives of the study were to (i) identify the breadth of conditions commonly encountered in primary care, (ii) create plausible combinations of conditions within synthetic patient profiles to maximise patient variability and (iii) evaluate prompting strategies and LLM behaviour when generating clinical correspondence and documentation associated with these profiles.

Methods and Data

A set of synthetic patient profiles was constructed to reflect the diversity of conditions encountered and complexity of patients in primary care. Open-source epidemiological statistics, clinical literature, and publicly available information on disease prevalence and comorbidities were used to identify common conditions managed in general practice, and typical patient characteristics. Twelve synthetic patient profiles represented varied demographic characteristics, combinations of conditions, and plausible diagnostic pathways.

Profiles were used as structured prompts to generate clinical documentation typically appended to general practice records following secondary care interactions, specifically requesting structures mapping to diagnostic letters, investigation reports, imaging summaries, referral correspondence, and administrative documentation. For each patient profile, multiple document types were requested, with an initial target of approximately 5-10 documents per patient. Prompting strategies were iteratively adjusted to improve the formatting, structure, and informational complexity.

Two open-source LLM systems (Claude.ai and ChatGPT) were used to generate 37 documents using multiple prompt variations for 12 patients. Outputs were reviewed qualitatively to assess their similarity to real-world clinical documentation, focusing on formatting, structural realism, informational completeness, and presence of extraneous or administrative content typically found in medical records.

Results

Patient profiles were generated to represent complex patients using multiple NHS services.

Table 1: Complex patient profiles for general practice.

Age / Sex	Conditions
32 F	Pregnancy (third trimester), Iron Deficiency Anaemia, Hypertension, Anxiety
72 F	Dementia, Hypertension, Osteoarthritis, Hearing Loss, Falls
15 M	Childhood obesity, ADHD, Depression, Anxiety, Violent outbursts
41 F	Psoriasis; Inflammatory arthritis (initially labelled OA); IBS; Hypothyroidism
52 F	Menopause; migraine; Anxiety/Depression; Urinary incontinence; Joint pain
46 M	Erectile dysfunction; Type 2 diabetes; Obesity; Depression; Sleep apnoea
29 F	Chronic pain; Anxiety; reflux; IBS; Orthostatic hypotension; Dysmenorrhea; Prolapse
33 M	ADHD (late diagnosis); Alcohol misuse; Anxiety/Depression; Smoking
79 F	CKD; Anaemia; Heart failure; Recurrent UTIs; Falls; Polypharmacy
44 F	Hypothyroidism; Depression; Chronic fatigue; Weight gain; Constipation
36 F	Chronic pain; IBS; Anxiety/Depression; Fibromyalgia
27 M	Orthostatic hypotension; ME/CFS; BPPV; Falls; Anxiety; IBS

The LLMs demonstrated strong performance in interpreting prompts correctly and generating clinically coherent narratives. Condition combinations and diagnostic pathways generated were typically plausible using appropriate medical terminology. Models were able to maintain longitudinal consistency when generating multiple documents associated with a single patient profile.

However, generated documents differed substantially from real clinical records, limiting their usefulness for evaluating healthcare NLP systems:

1) **Document structure and informational complexity**. Real medical records often contain large amounts of administrative and contextual information, partial data, formatting artifacts, and clinically irrelevant details. In contrast, LLM-generated documents tended to present information in a highly condensed and narrative form. For example, laboratory reports contained only the specific test values directly related to the prompted condition rather than a full panel of results. Imaging reports were frequently generated as short narrative summaries rather than structured reports. Administrative documents contained only minimal demographic information rather than extensive fields and metadata. Attempts to prompt the models to increase structural realism resulted in limited improvements. In some cases, additional formatting elements such as tables or checkboxes were added, but these often contained logically inconsistent information (e.g., mutually exclusive options marked simultaneously) or remained largely empty.

2) **Practical limitations** were also encountered during generation. The process of generating documents required substantial manual interaction, was slow and labour-intensive, making large-scale corpus generation impractical.

Conclusion

These findings suggest that while open-source LLMs can generate clinically coherent narratives, they currently struggle to reproduce the **structural complexity, formatting variation, and informational noise characteristic of real medical documentation**.

Study context

This study was conducted as part of exploratory research into methods for generating synthetic clinical datasets to support evaluation of healthcare NLP systems via Higher Education Innovation Funding. No real patient data were used at any stage of the project. All patient profiles and documents were entirely synthetic.

The work was undertaken within an academic research setting and did not involve human participants or identifiable clinical data. The study therefore did not require formal research ethics approval. The findings are intended to inform future methodological work on synthetic clinical data generation and evaluation frameworks for healthcare AI systems. As part of the same funding, consultations with GPs and patients are taking place to understand more about perceived benefits of the technology in the clinic to inform the future evaluation.

References

1. Johnson AE, Pollard TJ, Shen L, et al. (2016) MIMIC-III, a freely accessible critical care database. *Scientific data* 3(1). Nature Publishing Group: 1–9.
2. Wang Y, Wang L, Rastegar-Mojarad M, et al. (2018) Clinical information extraction applications: a literature review. *Journal of biomedical informatics* 77. Elsevier: 34–49.
3. Goncalves A, Ray P, Soper B, et al. (2020) Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology* 20(1): 108.
4. Tucker A, Wang Z, Rotalinti Y, et al. (2020) Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine* 3(1). Nature Publishing Group UK London: 147.

De-identifying patient records using a combination of regular expressions and large language models

Franz S Gruber[1,2], Stuart Dunbar[1], Matúš Falis[1,2], Samuel McInerney[1,2], Stephen Powell[1], Arlene Casey[1,2]

Middle authors listed alphabetically

[1] DataLoch, Usher Institute, University of Edinburgh, Edinburgh, UK

[2] Centre for Population Health Sciences, Usher Institute, University of Edinburgh, Edinburgh, UK

Introduction

Medical free-text data contain a vast amount of valuable information about patient journeys e.g. detailed radiology reports, or narrative summaries describing lifestyle factors or early symptoms long before a diagnosis is made. Most information in free-text data is not routinely transformed into structured, coded data and remains siloed, despite its potential to improve patient outcomes or inform public health policy. Because medical free-text data often contains sensitive personal information, organisations responsible for patient records are understandably cautious and frequently reluctant to make such data available for research [1].

Manual removal of sensitive information is costly, time-consuming, and labour-intensive. De-identification of clinical free-text data has been an ongoing subject for many years, yet the problem remains unsolved with relatively little research access to free-text data [2]. Large Language Models (LLMs) have shown promise in this area (e.g. [3]) but there has been no evaluation within Trusted Research Environments (TREs). Our project STAR-TRE ([4]) evaluates how LLMs and traditional NLP methods can work together to improve de-identification of clinical text. Here we present our work in progress, combining traditional pattern-based methods (regular expressions) with an LLM to achieve human-like removal of sensitive data in radiology reports and discharge summaries. Our tool is being developed for usage within TREs to facilitate governance decisions and potentially enable free-text data access to researchers.

Methods and Data

Data used for this work were sourced through DataLoch and represents a local extract of one year of radiology reports and discharge summary of patients utilizing the Lothian health board (south-east Scotland; total population of approximately 1 million).

We have previously double annotated a sample of 2,250 free-text reports [5] for each type (radiology and discharge). This dataset allowed us to create and evolve our in-house, python-based de-identification regular expression based tool consisting of a set of 19 regular expressions (e.g. postcodes, dates, or medical IDs). We are testing different prompting strategies but here we are focussing on few-shot prompting, to detect names (first name, last name) and dates. These entities are stored in the form of entity type, start of match, end of match, matched string – matching the output of the manual annotation tool brat [6].

A limitation we are currently working on is creating a gold standard set from our double annotated human files. At this point, results presented as part of this work stem from comparison to a single human annotator (1,250 files).

For evaluation we have adapted a system suggested by David Batista [7], which allows us to look at strict (i.e. same entity type AND same span match as human), exact (i.e. same span

match as human; independent of entity type), partial (some overlap of the match; independent of entity type) and type (some overlap of the match and same type) matches, when calculating precision, recall and f1-score. Here we only report F1-scores for strict matches comparing human annotation to either the pattern-based algorithm only or combined with LLM (Llama-3.1-70B [8]). Our pattern-based algorithm only detects patient names using structured name fields.

All our work has been performed locally within secured NHS computing infrastructure.

Results

Our de-identification tool is still in development, here we present performance metrics for names (first name, last name) and dates. We have compared the output of our de-identification tool to 1,250 human-annotated texts for each report type (radiology or discharge). Preliminary results are summarised in the table below:

Label	**F1-Score**	
Radiology	*Pattern-based*	*With LLM*
First name	0.15	0.98
Last name	<0.01	0.96
Dates	0.98	0.98
Discharge		
First name	0.44	0.95
Last name	0.39	0.89
Dates	0.71	0.90

Conclusion

Our initial results are encouraging and demonstrate the benefit of a hybrid de-identification approach for our purpose. One of the strengths of LLMs is their ability to consider linguistic context when identifying entities such as personal names. While rule-based algorithms can also account for context through increasingly complex patterns, doing so often requires substantial domain knowledge. For example, suppressing “John” as a patient name in “St. John’s Hospital” would require building numerous rules to capture variations, misspellings, and edge cases. An LLM, however, is more likely to understand that “John” in this context does not refer to a person but rather a location.

Outlook

The output of our de-identification tool will be displayed by a privacy risk dashboard (work in progress) which will be used by our data governance specialists. This dashboard will help to assess and audit projects requesting free-text data, with different amounts of granularity: What data will be accessed together with the free-text data? What type of free-text data is requested? What are the risks associated with making de-identified free-text data available to researchers? The dashboard should facilitate data governance decisions, ultimately to enable data access by researchers.

We will be piloting our tool and the dashboard internally, before opening it up to a wider audience. Our de-identification tool is open source, does not involve any model training, and can be deployed within moderate computing infrastructure (2x A100 GPUs). We are tailoring our tool to be used within other TREs.

Study context

This project was funded by DARE UK UKRI3005 (STAR-TRE project). A.C is funded by the Vivensa Foundation (PF2302\2).

This work uses data collected by the NHS as part of their provision of patient care and support. Following exceptional approval by NHS Lothian, the analysis was done within secured NHS computing infrastructure managed by the DataLoch service: a partnership of the University of Edinburgh and NHS Lothian. The data used for this project cannot be requested.

References

[1] Ford E et al., What is the patient re-identification risk from using de-identified clinical free text data for health research? AI Ethics, 2025

[2] Kovačević A et al., De-identification of clinical free text using natural language processing: A systematic review of current approaches. Artificial Intelligence in Medicine, 2024

[3] Wiest, I C et al., Deidentifying medical documents with local, privacy-preserving large language models: The LLM-Anonymizer. *NEJM AI*, 2025

[4] https://dareuk.org.uk/how-we-work/ongoing-activities/dare-uk-next-gen-catalysts/star-tre/

[5] Casey, A et al., Developing a Common Schema for De-identification of Personal Health Identifiers in EHRs across Scotland. HealTAC 2024

[6] Stenetorp P et al., brat: a Web-based Tool for NLP-Assisted Text. Association for Computational Linguistics, 2012 https://brat.nlplab.org/

[7] https://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/

[8] Aaron Grattafiori et al., The Llama 3 Herd of Models, arXiv, 2024 https://arxiv.org/abs/2407.21783

Metric-Dependent Optimisation of Clinical Prediction Models in Psychiatry

Frida Hæstrup Msc[1,2,3], Jakob G. Damgaard Msc[1,2,3], Sara Kolding Msc[1,2,3], Erik Perfalk MD PhD[1,2,3], Andreas A. Danielsen MD PhD[2,4], Søren D. Østergaard MD PhD[1,2]

[1]Department of Affective Disorders, Aarhus University Hospital, Denmark
[2]Department of Clinical Medicine, Aarhus University, Denmark
[3]Aarhus NLP group, Center for Computing Humanities, Aarhus University, Denmark
[4]Psychosis Research Group, Aarhus University Hospital, Aarhus, Denmark

Introduction

Clinical prediction models are emerging across a wide range of medical fields, often with the goal of diagnostic classification or predicting patient outcomes. Despite their initial promise, an increasing body of literature raises concerns about the real-world applicability of clinical prediction models and calls for validation of generalisability before such models are implemented in clinical practice [1–3]. How do we validate whether a 'good-performing' prediction model works well in a clinical context?

Training machine learning models often involves hyperparameter tuning, i.e., iterating over combinations of hyperparameters to find the 'best' combination. Usually, this tuning is performed to optimise the area under the receiver operating characteristic curve (AUROC). However, depending on the context in which the model is to be applied, different performance measures can be deemed more or less significant compared to others. As an example, [4] argue that ROC curves can be misleading and that the precision-recall curve (PRC) is more informative when working with imbalanced datasets. On a similar note, they highlight other approaches, such as the concentrated ROC curve (CROC), which explicitly emphasises performance in clinically relevant regions of the ROC space, e.g., at low false-positive rates [5].

In this study, we want to systematically investigate how the choice of performance metrics used during hyperparameter optimisation influences model behaviour, performance, and clinical usefulness in psychiatric prediction models.

Methods and Data

The study leverages a unique opportunity, namely a group of previously published prediction models of different clinical outcomes trained on data from the same population of patients from the Psychiatric Services of the Central Denmark Region [6]. The prediction models cover a range of different clinical outcomes, including prediction of type 2 diabetes [7], cardiovascular disease [8], schizophrenia or bipolar disorder [9], involuntary hospitalisation [10], physical restraint [11], and electroconvulsive therapy [12].

The data cohort encompasses routine clinical electronic health records (EHR) data from patients receiving treatment in the Psychiatric Services of the Central Denmark Region from January 1, 2013, to November 22, 2021. Predictors were derived from the EHRs, covering both structured data (diagnoses, medication, lab results, etc.) and unstructured free text data from clinical notes.

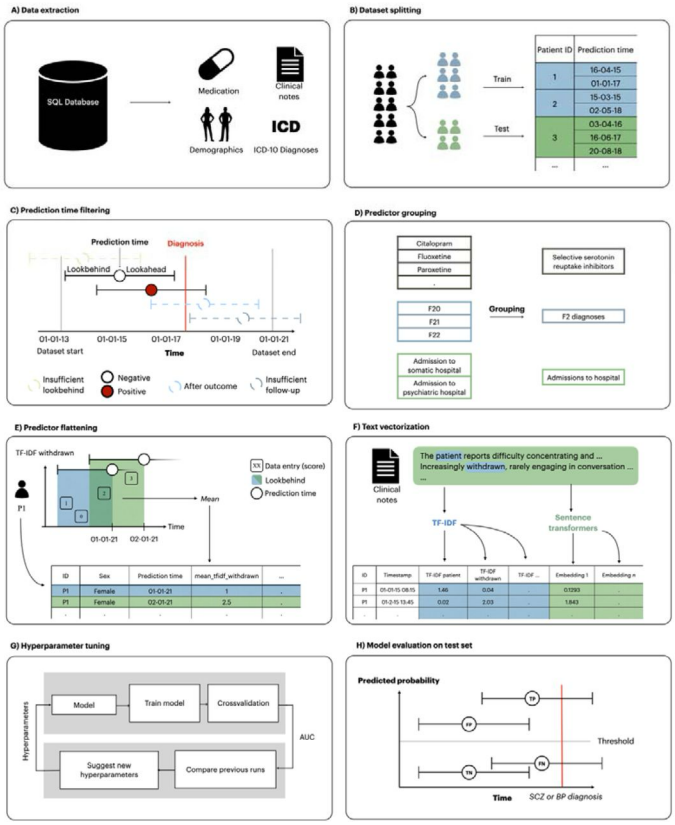


Figure 1. Overview of the process for extraction and transformation of dataset and the training and testing of models. A) Data as extracted from the EHRs. B) Data was split into a training and a test set. C) Prediction times occurring after September 22, 2021 and before January 1, 2015 were removed due to lack of follow-up/lookbehind in addition to prediction times preceded by diagnoses of psychotic or personality disorders. D) Linked predictors such as medication class and diagnostic groups were grouped together. E) Predictors for each prediction time were extracted by aggregating the variables within the lookbehind with an aggregation function. As a result, each row in the dataset represents a specific prediction time with a column for each predictor. F) Clinical notes were turned into vectors using TF-IDF models and sentence transformers. G) Models were trained and optimised on the training set using 5- fold cross-validation. Hyperparameters were tuned to optimise AUROC. H) The best candidate models were evaluated on the test set. Figure and description modified from [12].

Figure 1 provides an example of the original pipeline for data extraction and transformation and model training and testing. This figure illustrates the pipeline used for prediction of electroconvulsive therapy and is adapted from [12]. Although each of the included prediction models is accompanied by its own version of a similar pipeline figure, the overall training procedure is highly overlapping across the different outcomes.

We will follow the original model development pipelines described in the respective studies [7–12] with one modification, namely the hyperparameter optimisation procedure. For each of the outcomes, we will retrain a model using the original model class, data splits, and feature sets, but perform hyperparameter optimisation with respect to different performance metrics, such as the CROC or the PRC, rather than AUROC alone. All retrained models will be evaluated on the same held-out test set to isolate how the choice of optimisation metric influences discrimination and clinically relevant model behaviour. Evaluation will include predictive performance, robustness checks, and sensitivity analyses across the different optimisation regimes and clinical outcomes.

Results and Conclusion

This study is currently ongoing, and the results are not yet ready for publication. Preliminary results are expected to be ready for the conference.

Study context

The study is supported by grants to Søren D.Østergaard from the Lundbeck Foundation (grant No. R344-2020-1073), the Danish Cancer Society (grant No. R283-A16461), the Central Denmark Region Fund for Strengthening of Health Science (grant No. 1-36-72-4-20), the Danish Agency for Digitisation Investment Fund for New Technologies (grant No. 2020-6720), and Independent Research Fund Denmark (grant No. 4309-00028B).

The funders of the study had no role in design and conduct of this study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript or the decision to submit for publication.

The data from the study cannot be shared according to Danish law.
The code for all analyses will be available at:
https://github.com/Aarhus-Psychiatry-Research/psycop-common/tree/main

Søren D.Østergaard reported receiving grants from The Novo Nordisk Foundation (grant No. NNF20SA0062874), The Lundbeck Foundation (grant No. R358-2020-2341), and Independent Research Fund Denmark (grant No. 2096-00055B)); receiving the 2020 Lundbeck Foundation Young Investigator Prize; owning or having owned units of mutual funds with stock tickers DKIGI, IAIMWC, SPIC25KL, WEKAFKI, and DKIEUIXBNP; and owning or having owned units of exchange traded funds with stock tickers BATE, TRET, QDV5, QDVH, QDVE, SADM, IQQH, USPY, EXH2, 2B76, IS4S, OM3X, EUNL and SXRV outside the submitted work. No other disclosures were reported.

References

[1] Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605. https://doi.org/10.1136/bmj.b605.
[2] Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology 2016;69:245–7. https://doi.org/10.1016/j.jclinepi.2015.04.005.
[3] Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Medicine 2023;21:70. https://doi.org/10.1186/s12916-023-02779-w.
[4] Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE 2015;10:e0118432. https://doi.org/10.1371/journal.pone.0118432.
[5] Swamidass SJ, Azencott C-A, Daily K, Baldi P. A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. Bioinformatics 2010;26:1348–56. https://doi.org/10.1093/bioinformatics/btq140.
[6] Hansen L, Enevoldsen KC, Bernstorff M, Nielbo KL, Danielsen AA, Østergaard SD. The PSYchiatric clinical outcome prediction (PSYCOP) cohort: leveraging the potential of electronic health records in the treatment of mental disorders. Acta Neuropsychiatrica 2021;33:323–30. https://doi.org/10.1017/neu.2021.22.
[7] Bernstorff M, Hansen L, Enevoldsen K, Damgaard J, Hæstrup F, Perfalk E, et al. Development and validation of a machine learning model for prediction of type 2 diabetes in patients with mental illness. Acta Psychiatrica Scandinavica 2024. https://doi.org/10.1111/acps.13687.
[8] Bernstorff M, Hansen L, Olesen KKW, Danielsen AA, Østergaard SD. Predicting cardiovascular disease in patients with mental illness using machine learning. European Psychiatry 2025;68:e12. https://doi.org/10.1192/j.eurpsy.2025.1.

[9] Hansen L, Bernstorff M, Enevoldsen K, Kolding S, Damgaard JG, Perfalk E, et al. Predicting Diagnostic Progression to Schizophrenia or Bipolar Disorder via Machine Learning. JAMA Psychiatry 2025;82:459–69. https://doi.org/10.1001/jamapsychiatry.2024.4702.
[10] Perfalk E, Damgaard JG, Bernstorff M, Hansen L, Danielsen AA, Østergaard SD. Predicting involuntary admission following inpatient psychiatric treatment using machine learning trained on electronic health record data. Psychological Medicine 2024;54:4348–61. https://doi.org/10.1017/S0033291724002642.
[11] Kolding S, Damgaard JG, Bernstorff M, Hansen L, Østergaard SD, Danielsen AA. Development and Evaluation of Machine Learning Models to Predict Mechanical Restraint and Related Coercive Measures in Hospital Psychiatry 2025:2025.12.15.25342272. https://doi.org/10.64898/2025.12.15.25342272.
[12] Hansen L, Damgaard JG, Lundin RM, Danielsen AA, Østergaard SD. Predicting the need for electroconvulsive therapy via machine learning trained on electronic health record data. Acta Neuropsychiatrica 2026:1–23. https://doi.org/10.1017/neu.2026.10063.

Learning from Tragedy: Structuring Complex Narrative Evidence in Health Systems with Ontology-Guided Hybrid NLP

Paul Howarth[1]
[1]Akumen Ltd, Bude, United Kingdom

Study Context

Akumen commissioned by Cornwall Council Public Health, analysed sensitive coroner investigation materials under appropriate governance. This work supports research into constraint-driven inference and ontology-guided narrative analytics to detect low-prevalence, high-impact signals in sparse narrative environments. The methodology was developed to strengthen how narrative evidence informs system learning and prevention.

Introduction

Signals relating to risk, failure, and vulnerability in health systems are embedded in narrative records rather than structured datasets. Investigative reports, case reviews, and experiential accounts describe events, contextual conditions, service interactions, and systemic factors shaping outcomes. However, insight is distributed across long, heterogeneous documents that resist systematic analysis, leaving much clinically relevant information locked within unstructured narrative material (Kreimeyer et al., 2017).
Narrative investigative records represent a high-stakes evidence source in which small signals may have important implications for prevention and system learning. Yet extracting those signals is difficult because relationships between events, actors, environments, and decisions are embedded within the structure and sequence of narrative accounts rather than structured data fields.
Most healthcare text analytics relies on statistical natural language processing methods that identify patterns through probabilistic associations or classification models. However, these approaches can struggle when analysis requires preservation of meaning across complex narratives and the relationships they describe (Rajkomar, Dean and Kohane, 2019; Usuyama et al., 2024).

This challenge is particularly evident in the analysis of suicide cases and related investigative documentation. Coroners' reports contain detailed accounts of personal histories, service interactions, environmental conditions, and contributing factors surrounding deaths. Their volume and complexity make systematic analysis difficult using traditional manual approaches. In regions of Southwest England, approximately one person dies by suicide every five and a half days.

Methods and Data

The methodology structures complex narrative evidence using ontology guided, meaning based natural language analysis. Narrative documents are treated as structured descriptions of events, actors, environments, and experiences represented through explicit semantic frameworks.
Akumen uses large language models within a governed human-led process to help identify themes and extract key narrative phrases from unstructured data. Agentic AI workflows refine and map these phrases to structured ontologies, improving thematic consistency from roughly 60% out-of-the-box accuracy to around 80%. Final interpretive decisions remain with human analysts, ensuring contextual judgement, traceability, and evidential integrity while transforming unstructured narrative into structured analytical evidence.

Akumen have developed a mental health ontology which treats mental health as a complete human system rather than a collection of isolated symptoms or themes. Drawing on the TRIZ Law of System Completeness, we developed TRACES: a systems-based framework in which mental health is understood through the interconnected relationship between Thought, Result, Automatic Thought, Character and Conduct, Emotions, and Sensing.
A layered ontology captures key dimensions of investigative narratives including service interaction, psychological and social context, environmental conditions, temporal sequence, and wider system influences. Narrative material is processed through automated semantic extraction aligned with these ontology structures.
The analytical process combines three components. Deductive ontology structures organise narrative evidence within a consistent semantic framework, while inductive thematic analysis allows additional patterns to emerge from the narratives. A human interpretive layer introduces contextual discernment where required, ensuring interpretation remains grounded in the source material.
This hybrid design prioritises transparency and traceability, with extracted narrative elements linked to their originating text. As materials are structured, they form a cumulative evidence base enabling cross case comparison and multiple analytical interrogations.

The methodology was demonstrated through analysis of seventy-nine coroner investigation bundles comprising approximately 5,600 pages of narrative documentation describing deaths involving complex personal, clinical, and systemic factors. Coroners' investigations and related mortality review processes represent an important mechanism through which health systems attempt to identify contributory factors and prevent future deaths (Pudney and Grech, 2016).
Structuring this material enables patterns of vulnerability, service interaction breakdown, and missed opportunities for intervention to be examined across cases in ways difficult to achieve through manual review or purely statistical text analysis.
Processing was undertaken under a formal data sharing and data processing agreement between the client acting as Data Controller and Akumen acting as Data Processor in accordance with UK GDPR and relevant public sector information governance requirements.

Results

The analysis identified fifty-three system level themes derived through combined deductive and inductive thematic analysis. Deductive themes were informed by research into Regulation 28 Reports to Prevent Future Deaths, while inductive analysis identified additional themes emerging from investigative narratives.
Three new themes were incorporated into Akumen's Assisted Mental Health Ontology, expanding the ontology from 92 to 95 themes. Across the dataset, more than 10,000 narrative phrases were extracted and mapped to thematic structures.
These findings demonstrate the feasibility of structuring large volumes of investigative narrative material and establishing a basis for systematic cross case analysis of coronial records.

Conclusion

Health systems generate extensive narrative documentation containing important but difficult to analyse insight. This study demonstrates a methodology for structuring such material using ontology guided semantic analysis supported by hybrid human and machine processes.
Applied to a pilot dataset of coroner investigation bundles, the approach shows that complex narrative evidence can be transformed into structured analytical datasets while preserving meaning, context, and traceability.

More broadly, the methodology indicates that narrative records represent an untapped intelligence source within health systems. Structuring these materials enables real world narrative evidence to be processed within transparent analytical platforms supporting system learning and improvement.

References

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S., Forshee, R., Walderhaug, M. and Botsis, T. (2017)
Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review.
Journal of Biomedical Informatics, 73, pp. 14–29.

Rajkomar, A., Dean, J. and Kohane, I. (2019)
Machine learning in medicine.
New England Journal of Medicine, 380(14), pp. 1347–1358.

Usuyama, N., Wong, A., Zhang, Y., Naumann, T. and Poon, H. (2024)
Biomedical natural language processing in the era of large language models.
Annual Review of Biomedical Data Science, 7.

Pudney, E. and Grech, E. (2016)
Systematic analysis of coroners' inquest data to inform patient safety improvement.
BMJ Open, 6(2).

Assessing Certainty of Diagnoses in Clinical Text

Arooj Hussain[1], Warren Del-Pinto[1], Meghna Jani[1,2], William G. Dixon[1,2], Goran Nenadic[1]
[1] The University of Manchester, Manchester, United Kingdom
[2] Northern Care Alliance – NHS Foundation Trust, Salford, United Kingdom

Introduction

Most clinical Natural Language Processing (NLP) models extract the mentions of diagnoses without considering the associated 'certainty qualifiers' that may alter their applicability (see examples in Figure. 1). With models extracting all the mentions of diagnoses as confirmed, there is a high risk that the research conducted on the extracted data, such as for public health studies, is actually based on inflated counts which could result in inaccurate results. While some approaches include certainty detection [1-3] in this study we evaluated various NLP paradigms and models on the task of recognising the certainty qualifiers contextualising diagnoses, including implicitly mentioned qualifiers.

She denies having muscle aches
MI has been ruled out
Chest pain due to GERD rather than MI
(A)

Probable norovirus infection
Suspected appendicitis
Findings consistent with Pneumonia
(B)

Figure. 1. Examples of (A) negated and (B) uncertain certainty qualifiers w.r.t diagnoses

The study objectives were to:

1. Explore the accuracy of rule-based and transformer-based NLP models in extracting the mentions of certainty qualifiers along with diagnoses;
2. Investigate the effects of different strategies in prompt engineering and fine-tuning;
3. Conduct an error analysis to identify similarities and differences in misclassified cases across different models.

Methods and Data

The Clinical Assertion Data (CAD) [4] dataset used for this study has been developed from MIMIC III [5]. It has 5000 diagnosis entity mentions with corresponding mention-level certainty qualifiers categorised as confirmed, uncertain or negated. Around 69% of the diagnoses have been labelled as confirmed, ~24% as negated and ~7% as uncertain. Keeping the class distribution same across the split, we used 15% of the data as a test set and used the rest of it for training and development.

We compared six models including at least one each from the major NLP paradigms: rule-based (*MedspaCy* [6]*)* small-scale transformer-based (*BioClinicalBERT* [7]) and large-scale transformer-based language models. The latter included two general purpose LLMs, *Gemma-2-8b* [8]*, Llama-3.1-7b* [9] and two clinical LLMs, *Asclepius-7b* [10] *and MedAlpaca-7b* [11].

The four LLMs were tested in both zero- and few-shot [12] settings. BioClinicalBERT as well as Gemma and Llama were fine-tuned using the training and validation sets, with BioClinicalBERT being finetuned fully whereas Gemma and Llama only partially using LORA [13]. MedspaCy was tuned to the task by adding extra rules based on error evaluation on the training and development set.

Results

The best overall performance was achieved by fine-tuned Gemma followed by BioClinicalBERT. Prompted general-purpose LLMs showed mixed results with Few-shot prompting with Gemma achieving relatively strong performance compared to Few-shot Llama. Prompted clinical LLMs showed more limited performance overall whereas rule-based model MedspaCy performed better on the uncertain class compared to the prompted LLMs. For comparison, we also calculated scores for a

random classifier and a model that predicts all cases as confirmed. See Table 1 for the best-performing variant for each model, with column-wise best scores highlighted.

Table 1. Models' results. (key to column names – C: Confirmed, N: Negated, U: Uncertain)

	Model	Macro F1 score	Micro F1 score	Class-wise Precision			Class-wise Recall		
				C	N	U	C	N	U
Common baselines	Random classifier	0.2502	0.2840	0.6600	0.2300	0.0600	0.2700	0.3100	0.3200
	All Confirmed (majority class)	0.2700	0.6900	0.6900	0.0000	0.0000	1.0000	0.0000	0.0000
Rule-based methods	MedspaCy	0.6838	0.8647	0.8916	0.8304	0.5000	0.9316	0.8712	0.2041
	MedspaCy ++	0.7648	0.8662	0.9356	0.8354	0.4697	0.8996	0.8405	0.6327
BERT-based model	BioClinicalBERT	0.8513	0.9324	0.9436	0.9620	0.6977	0.9658	0.9325	0.6122
LLMs prompted	Few_shot_ent_Gemma	0.7972	0.8838	0.9518	0.9613	0.4157	0.8868	0.9141	0.7551
	Few_shot_Llama	0.5638	0.5632	0.9489	**0.9655**	0.1459	0.4765	0.6871	**0.9796**
Clinical LLMs prompted	Zero_shot_Asclepius	0.5487	0.6500	0.8447	0.7255	0.1313	0.6624	0.6810	0.4286
	Zero_shot_MedAlpaca	0.3252	0.5926	0.6850	0.3000	0.1600	0.8269	0.0736	0.0816
LLMs finetuned	Gemma_org_ent	**0.9172**	**0.9574**	**0.9762**	0.9518	**0.8039**	**0.9658**	**0.9693**	0.8367
	Llama_simp_ent_role	0.8512	0.9309	0.9569	0.9458	0.6400	0.9487	0.9632	0.6531

Most models exhibited largely distinct error patterns, with especially unique errors for Few_shot_Llama, Zero_shot_Asclepius, and the rule-based approaches. Most overlap in errors was observed between Few_shot_Llama and Zero_shot_Asclepius, followed by MedspaCy and MedspaCy++. Manual error analysis revealed some issues like models extending the scope of the qualifier beyond the actual entity, trouble recognising implicit qualifiers in complicated sentence structure as well as some discrepancies in the gold standard.

Conclusion

Our results revealed that fine-tuned *Gemma* performed the best on detecting negation and uncertainty cues, with a comparable performance from *BioClinicalBERT*.

Study context

The CAD dataset is available for anyone to use once they have signed the Agreement of Use with PhysioNet (a repository of medical research data managed by the Massachusetts Institute of Technology) for access to MIMIC datasets. There are no conflicts of interest.

References

[1] H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman, 'ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports', *J. Biomed. Inform.*, vol. 42, no. 5, pp. 839–851, Oct. 2009, doi: 10.1016/j.jbi.2009.05.002.

[2] L. Chen, 'Attention-Based Deep Learning System for Negation and Assertion Detection in Clinical Notes', *Int. J. Artif. Intell. Appl.*, vol. 10, no. 01, pp. 1–9, Jan. 2019, doi: 10.5121/ijaia.2019.10101.

[3] S. Wang *et al.*, 'Trustworthy assertion classification through prompting', *J. Biomed. Inform.*, vol. 132, p. 104139, Aug. 2022, doi: 10.1016/j.jbi.2022.104139.

[4] B. Van Aken, I. Trajanovska, A. Siu, M. Mayrdorfer, K. Budde, and A. Loeser, 'Assertion Detection in Clinical Notes: Medical Language Models to the Rescue?', in *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, Online: Association for Computational Linguistics, 2021, pp. 35–40. doi: 10.18653/v1/2021.nlpmc-1.5.

[5] A. E. W. Johnson *et al.*, 'MIMIC-III, a freely accessible critical care database', *Sci. Data*, vol. 3, no. 1, p. 160035, May 2016, doi: 10.1038/sdata.2016.35.

[6] H. Eyre *et al.*, 'Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python'.

[7] E. Alsentzer *et al.*, 'Publicly Available Clinical BERT Embeddings', 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/

[8] G. Team *et al.*, 'Gemma 2: Improving Open Language Models at a Practical Size', Oct. 02, 2024, *arXiv*: arXiv:2408.00118. doi: 10.48550/arXiv.2408.00118.

[9] L. Team and A. @ Meta, 'The Llama 3 Herd of Models', 2024.

[10] S. Kweon *et al.*, 'Publicly Shareable Clinical Large Language Model Built on Synthetic Clinical Notes', Sep. 2023, [Online]. Available: http://arxiv.org/abs/2309.00237

[11] T. Han *et al.*, 'MedAlpaca -- An Open-Source Collection of Medical Conversational AI Models and Training Data', Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.08247

[12] T. B. Brown *et al.*, 'Language Models are Few-Shot Learners', May 2020, [Online]. Available: http://arxiv.org/abs/2005.14165

[13] E. J. Hu *et al.*, 'LoRA: Low-Rank Adaptation of Large Language Models', Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.09685

AI-enhanced dementia prevention: precision risk reduction using large language models

Chloe Hutton[1], Chris Fox[1], Carol Brayne[2], Hang Dong[1], David Llewellyn[1]
[1] University of Exeter, Exeter, United Kingdom
[2] University of Cambridge, Cambridge, United Kingdom

Introduction

Dementia is a global public health challenge and a leading cause of dependence and disability (1). It is estimated that 78 million people will have dementia by 2030, rising to 139 million in 2050 (1). Large costs have been associated with the disease, placing a strain on services and communities (2).

There is increasing evidence that addressing modifiable risk factors may prevent or delay nearly half of dementia cases (3). Fourteen potentially modifiable risk factors have been identified irrespective of a person's apolipoprotein E (APOE) genetic status, with some of these associated more prominently with particular periods of life (3). Population-based policy can help address these risk factors (3). Prior studies have shown a decline in the rate of dementia cases in higher-income countries, which may be associated with a reduction in risk factors due to country-level policies to improve education and reduce smoking rates (4). Risk factors vary between individuals, and a targeted multicomponent risk reduction approach may be beneficial (3). Identifying effective strategies to communicate modifiable risk factors across different cultures and settings is a research priority globally (5). In England, a shift from sickness to prevention is one key shift in its new ten-year health plan (6).

Another focus in England's ten-year health plan is a shift from analogue to digital (6). Machine learning (ML) could improve dementia prevention (7). While most ML methods tested so far have not been found to outperform traditional analytical risk profiling methods (8), they may be more cost-effective and can cope with a greater number of variables than traditional approaches (9).

Generative Artificial Intelligence (AI) is a section of unsupervised machine learning that creates original content from user prompts using generative models based on deep learning models (10, 11). Generative AI has the potential to personalise risk reduction strategies as well as the communication of these to the individual. However, ethical and practical considerations need to be addressed (12), a principle highlighted by NHS England (13).

This project aims to explore how large language models (LLMs), a form of generative AI, can improve the identification of dementia risks and help create personalised strategies to reduce those risks, focusing on factors that can be modified. This will be achieved through the following objectives:

1. Synthesise the existing literature on the use of generative AI healthcare applications for dementia
2. Build a generative AI prototype for personalised dementia risk profiling and reduction strategies with support from expert consensus for the practical and ethical guidelines
3. Evaluate the generative AI prototype

Methods and Data

The following methods are proposed for the project:

Systematic Review

The systematic review will identify and evaluate existing generative AI models applied to dementia. Its application will be evaluated, as well as identifying any challenges or opportunities that arise. The findings from the systematic review will inform the initial scoping for the LLM requirements as well as the design of the Modified Delphi Consensus Study.

The review has been registered on PROSPERO (CRD420261306781) which provides more information on the protocol. The review will be documented following PRISMA reporting guidelines (14).

Modified Delphi consensus study

A panel of experts will be engaged to achieve consensus on the design considerations for the prototype. The results will be used to develop the LLM requirements. A protocol will be finalised and published prior to commencing and the study design will follow the ACCORD reporting guidelines for consensus methods (15).

Like other Delphi studies (16, 17), a steering group will be established which will help steer the study process and finalise various study decisions and issues. This will include agreeing on the consensus threshold, criteria for dropping or refining statements, identifying panellists to take part in the study, organisations to approach, providing input into an initial 'round 0' workshop and reviewing results between rounds. Ideally, the steering group would include members of the project's supervisory team who can provide expertise from a clinical and technical perspective, and public patient involvement and engagement (PPIE) representation.

Panel members will be representatives from patient advocacy groups or experienced caregivers, PPIE, clinicians (psychiatrists, neurologists, geriatricians), researchers in dementia risk and prevention, and those with healthcare technology expertise. Panellists will be recruited using a purposeful approach through the steering group identifying individuals with relevant expertise and experience. Additionally, an open advert will be shared with relevant organisations including UK Brain Health Coalition, DEMON network and The European Brain Research Area (EBRA) Consortium, Alzheimer's Society and Dementia UK. This may help to improve the representativeness of the panel. An initial screening process will be developed to ensure that the prospective participants had sufficient experience or expertise to contribute.

The study will aim to recruit around 70 panellists. Due to the required heterogeneity of the panel due to the topic, a panel size between 60 and 80 may be sufficient to represent different stakeholder groups and help stabilise results (18). While there is no set number of participants for a Delphi study, most studies recruit into the double digits and would aim to account for a potential dropout rate around 19% (19). A round zero workshop is planned to mitigate some of the dropout rate by setting expectations of involvement and promoting engagement.

An initial online workshop (round zero) will be organised to provide participants with further context of the study and to gather feedback to inform the survey content for round 1. Up to three Delphi rounds have been planned within the timeline. Three rounds are deemed optimal to retain panel engagement and allow for feedback (19), and would be feasible within the time constraints of the project.

The timeframe for responses from participants will be two weeks. A further two weeks will be used to evaluate the responses and refine the survey for the following round. The planned timeline would be shared at the online workshop and with any recruitment information circulated.

The initial online workshop will focus on gathering feedback on how LLMs could be designed appropriately to support dementia prevention and risk profiling workflows. A previous Delphi study highlighted the importance of evaluating the implementation of LLMs, not just testing their potential (20). A recent systematic review highlighted design limitations of LLMs for patient care which included a lack of optimisation for medical use, misunderstanding medical

information and terms, lack of context on the training data used, explainability of black-box algorithms and concerns over the lack of self-validation of models (21). Open questions related to design requirements will be needed to ensure a dementia prevention LLM application would be acceptable for patients and their family and/or caregivers, and usable for clinicians. Initial questions might be:

- What should a computer program be able to do (and not do) in dementia prevention and risk profiling?
- Where would a dementia risk and prevention tool fit into current dementia screening and prevention workflows?
- What would make a dementia risk and prevention tool practical to use?
- What information should a dementia risk and prevention tool provide to clinicians to support decision-making?
- What evidence would you need to see before trusting a computer program for dementia risk profiling?
- What would make patients feel confident that a computer program is providing trustworthy information?

The final design will be confirmed by the steering group guiding the study and piloted to consider appropriate wording and whether framing the questions with scenarios might be appropriate. Additionally, a key terms glossary to support any layperson involvement will be provided. It is likely that a five-point Likert scale or similar will be used with an additional option for 'don't know' to allow for non-response if someone on the panel feels unable to respond to a particular statement. A free text field would be provided at the end of the survey for panel members to include any additional information or feedback. The survey will be set up and distributed in a survey design platform such as Qualtrics or similar (22).

Custom LLM development

A custom large language model will be developed using an open-source base model and the feedback from the Modified Delphi Consensus study. It will be finetuned with domain-specific training data and feature a bespoke interface. It will analyse complex clinical features and patient-generated narratives to identify modifiable risk factors and generate personalised risk reduction strategies and recommendations.

The latest stable version of Meta Llama has been planned to use as the foundation model. Currently, this is Llama 4. Llama is a foundational model available under a bespoke commercial license which provides a royalty-free limited license to copy and modify subject to attribution and an acceptable use policy (23, 24). It can be hosted locally, run and provides the developer with control over who can access the inputs and outputs (25). This is important to comply with the End User Licence (EUL) Agreement for the proposed data source (26). A base model which is freely accessible and provides control over the way any data is stored is likely to be required for the prototype's design considerations as well (21). Using a local copy of the base model will avert issues arising from unplanned updates.

The English Longitudinal Study of Ageing (ELSA) has been identified as a potential data source to use for the study (27). The longitudinal study began in 2002 and focuses on economical, health, wellbeing and social contexts in the English population aged 50 and older with follow-up interviews every two years, described as waves, to track changes over time (28). Waves 1-11 can be downloaded through the UK Data Service, which the University of Exeter can access as part of the UK Access Management Federation (UKAMF) (29). It is a safeguarded data collection which can be downloaded by accepting the EUL agreement. This data source has been selected due to the variety of data it provides relating to many of the potentially modifiable risk factors for dementia as well as its availability and EUL agreement ensuring that it is acceptable to use and available within the required timeframe.

ELSA provides a weighting strategy to adjust for variations in selection probability and non-response for both longitudinal and cross-section analysis of waves based on analysing demographic variables against household population in England, excluding those living in institutions (28). For longitudinal analysis, this is available for participants who have remained in the study in every wave or through waves 4 to 11. This could be a preprocessing mitigation to create a more representative training dataset for the prototype (30) where the weights have already been calibrated by the ELSA team.

Model evaluation

The prototype will be assessed based on its acceptability, feasibility, and initial performance, specifically regarding its usability, capacity to identify risks and convey effective risk reduction strategies. Its performance will be compared to alternative methods and will incorporate feedback from healthcare professionals and the public.

Results

This project plans to obtain results through the following:

Systematic review findings. The systematic review will synthesise existing direct applications of generative AI in dementia, as well as considering the challenges and opportunities addressed by these studies.

Delphi consensus study. The Delphi study will help form a consensus on the design considerations of an LLM prototype designed for dementia risk reduction from a multidisciplinary panel of experts. The results of individual rounds will be reported using the consensus threshold, agreement gradings and analysis of any free text comments within the survey. Any revisions or dropped statements would be reported. The results will inform the LLM build.

LLM evaluation. The final model will gain an initial evaluation through PPIE and clinician feedback considering its usability and acceptability. The technical performance and bias will be tested and benchmarked with an established dementia risk assessment.

Conclusion

This project aligns well with international and national efforts to identify approaches to mitigate personalised dementia risk though individualised prevention strategies. Additionally, it aligns with the 10 Year health plan for England to shift from analogue to digital (6). The methodologies selected are designed to produce an informed prototype through prior research findings and stakeholder engagement and could form the basis of further research into designing and implementing this type of tool into the dementia care pathway.

Study context

Ethics is not required for the systematic review and Delphi consensus study due to the study design. Initial feedback on the LLM prototype is planned through PPIE and clinician feedback. The university ethics team will be consulted to check whether approval is required for this aspect of the project.

This PhD project is funded by the NIHR Three Schools Dementia programme (reference: 5100) (31).

PPIE is planned for the Delphi study and LLM build. PPIE representation is planned for joining the Delphi study steering group, as well as providing feedback on the LLM development and initial evaluation. Delphi panel members will be representatives from patient advocacy groups or experienced caregivers, PPIE, clinicians (psychiatrists, neurologists, geriatricians),

researchers in dementia risk and prevention, and those with healthcare technology expertise. A full list of organisations to contact with an open advert to take part will be available in the protocol once finalised.

The systematic review is registered on PROSPERO (CRD420261306781) and the Delphi study protocol will be available on Open Science Framework once agreed by the steering group. The planned dataset to finetune the LLM, ELSA, is available through the UK Data Service for institutions part of the UK Access Management Federation (UKAMF) (29). The LLM build and evaluation method will be made available as well.

The supervisory team for this project is led by Professor David Llewellyn, Professor of Clinical Epidemiology and Digital Health at the University of Exeter Medical School. The rest of the supervisory team are Professor Chris Fox, Professor of Clinical Psychiatry and NIHR Senior Investigator at the University of Exeter Medical School, Professor Carol Brayne, Professor Emeritus and Senior Visiting Fellow in the Department of Psychiatry at the University of Cambridge, and Dr Hang Dong, Lecturer in Computer Science at the University of Exeter.

References

1. World Health Organization. Global status report on the public health response to dementia. Licence: CC BY-NC-SA 3.0 IGO. Geneva: World Health Organization; 2021.
2. Wimo A, Seeher K, Cataldi R, Cyhlarova E, Dielemann JL, Frisell O, et al. The worldwide costs of dementia in 2019. Alzheimer's & Dementia. 2023;19(7):2865-73.
3. Livingston G, Huntley J, Liu KY, Costafreda SG, Selbæk G, Alladi S, et al. Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission. The Lancet. 2024;404(10452):572-628.
4. Mukadam N, Wolters FJ, Walsh S, Wallace L, Brayne C, Matthews FE, et al. Changes in prevalence and incidence of dementia and risk factors for dementia: an analysis from cohort studies. The Lancet Public Health. 2024;9(7):e443-e60.
5. Shah H, Albanese E, Duggan C, Rudan I, Langa KM, Carrillo MC, et al. Research priorities to reduce the global burden of dementia by 2025. The Lancet Neurology. 2016;15(12):1285-94.
6. Department of Health and Social Care. Fit for the future: 10 year health plan for England. In: Department of Health and Social Care, editor. London: Crown copyright; 2025.
7. Bucholc M, James C, Khleifat AA, Badhwar A, Clarke N, Dehsarvi A, et al. Artificial intelligence for dementia research methods optimization. Alzheimer's & Dementia. 2023;19(12):5934-51.
8. Brain J, Kafadar AH, Errington L, Kirkley R, Tang EYH, Akyea RK, et al. What's New in Dementia Risk Prediction Modelling? An Updated Systematic Review. Dementia and Geriatric Cognitive Disorders Extra. 2024;14(1):49-74.
9. Newby D, Orgeta V, Marshall CR, Lourida I, Albertyn CP, Tamburin S, et al. Artificial intelligence for dementia prevention. Alzheimer's & Dementia. 2023;19(12):5952-69.
10. Fui-Hoon Nah F, Zheng R, Cai J, Siau K, Chen L. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. Journal of Information Technology Case and Application Research. 2023;25(3):277-304.
11. Stryker C, Scapicchio M. What is generative AI? [Place of publication not identified]: IBM; 2024 [cited 2025 January 26]. Available from: https://www.ibm.com/think/topics/generative-ai.
12. Ning Y, Teixayavong S, Shang Y, Savulescu J, Nagaraj V, Miao D, et al. Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. Lancet Digital Health. 2024;6(11):e848-e56.
13. NHSX. A Buyer's Guide to AI in Health and Care 2020 [cited 2025 August 26]. Available from: https://digital.nhs.uk/services/ai-knowledge-repository/develop-ai/a-buyers-guide-to-ai-in-health-and-care.

14. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
15. Gattrell WT, Logullo P, van Zuuren EJ, Price A, Hughes EL, Blazey P, et al. ACCORD (ACcurate COnsensus Reporting Document): A reporting guideline for consensus methods in biomedicine developed via a modified Delphi. PLOS Medicine. 2024;21(1):e1004326.
16. Lazarus JV, Romero D, Kopka CJ, Karim SA, Abu-Raddad LJ, Almeida G, et al. A multinational Delphi consensus to end the COVID-19 public health threat. Nature. 2022;611(7935):332-45.
17. Demnitz-King H, Banerjee S, Cooper C, Kenten C, Phillips R, Zabihi S, et al. The Nottingham consensus on dementia risk reduction policy: recommendations from a modified Delphi process. Nature Reviews Neurology. 2026;22(2):123-35.
18. Manyara AM, Purvis A, Ciani O, Collins GS, Taylor RS. Sample size in multistakeholder Delphi surveys: at what minimum sample size do replicability of results stabilize? Journal of Clinical Epidemiology. 2024;174:111485.
19. Schifano J, Niederberger M. How Delphi studies in the health sciences find consensus: a scoping review. Syst Rev. 2025;14(1):14.
20. Denecke K, May R, Rivera Romero O. Potential of Large Language Models in Health Care: Delphi Study. J Med Internet Res. 2024;26:e52399.
21. Busch F, Hoffmann L, Rueger C, van Dijk EHC, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Communications Medicine. 2025;5(1):26.
22. Qualtrics. Qualtrics 2026 [cited 2026 February 19]. Available from: https://www.qualtrics.com/.
23. Meta. llama 4 community license agreement 2025 [cited 2026 February 20]. Available from: https://github.com/meta-llama/llama-models/blob/main/models/llama4/LICENSE.
24. Meta. Llama 4 acceptable use policy 2025 [cited 2026 February 20]. Available from: https://github.com/meta-llama/llama-models/blob/main/models/llama4/USE_POLICY.md.
25. Meta. Troubleshooting & FAQ n.d. [cited 2026 February 20]. Available from: https://www.llama.com/faq/.
26. UK Data Service. End user license agreement 2024 [cited 2026 February 20]. Available from: https://ukdataservice.ac.uk/app/uploads/cd137-enduserlicence.pdf.
27. Banks J, Batty GD, Breedvelt J, Coughlin K, Crawford R, Marmot M, et al. English Longitudinal Study of Ageing: Waves 0-11, 1998-2024. 49th Edition ed: UK Data Service; 2026.
28. Lloyd L, Taylor K, Tsantani M, Englefield L, Williams R, Allen J. The dynamics of ageing: The 2023/2024 English Longitudinal Study of Ageing (Wave 11) Technical Report. 2025.
29. English Longitudinal Study of Ageing. Accessing ELSA data 2024 [cited 2026 February 20]. Available from: https://www.elsa-project.ac.uk/accessing-elsa-data.
30. Gallegos IO, Rossi RA, Barrow J, Tanjim MM, Kim S, Dernoncourt F, et al. Bias and Fairness in Large Language Models: A Survey. Computational Linguistics. 2024;50(3):1097-179.
31. NIHR School for Social Care Research. NIHR Three Schools' Dementia Research Programme 2026 [cited 2026 March 13]. Available from: https://sscr.nihr.ac.uk/dementia-programme/.

PAIR-SUM: Summarisation of MIMIC hypertension discharge notes

Yunsoo Kim[1,2], Xiaolei Diao[2], Chao Xu[2], Sandosh Padmanabhan[2], and Honghan Wu[1,2]

[1]University College London, London, UK
[2]University of Glasgow, Glasgow, UK

1 Introduction

Electronic health records (EHRs) contain detailed narratives documenting patient history, diagnosis, treatment, and follow-up plans [1, 2, 3, 4]. Discharge notes are a critical component of EHRs, providing clinicians with an overview of a patient's hospital stay and supporting continuity of care [5]. However, discharge notes are often lengthy and unstructured, which hinders rapid interpretation in time-constrained clinical settings [6, 7, 8].

Hypertension is a common comorbidity among intensive care patients in the MIMIC database, with management typically documented in discharge summaries that include blood pressure trajectories, antihypertensive regimens, and follow-up recommendations [9]. Most existing benchmarks, however, are phenotype-agnostic and often use discharge instructions as the reference summary rather than a coherent summary written specifically for the case [10].

To address this gap, **PAIR-SUM** is introduced as a curated dataset of hypertension-focused discharge note summaries derived from MIMIC. Each discharge note is paired with a specialist-written summary by a hypertension expert. This dataset enables evaluation of multiple medical large language models (LLMs) and a general domain models to determine whether biomedical pretraining improves the capture of clinically relevant information in hypertension-related discharge notes.

2 Methods

2.1 Dataset and Annotation

Twenty-seven discharge notes related to hypertension were curated from the MIMIC clinical database [5]. Each note was paired with a gold-standard summary written by a hypertension specialist. The summaries are organized into five clinically relevant sections: *demographics*, *medications*, *brief hospital course*, *vital signs*, and *laboratory testing results*. These sections capture the essential information required for hypertension management and follow-up care.

Although limited in scale, PAIR-SUM provides high-quality expert annotations and serves as a focused benchmark for disease-specific clinical summarisation.

	Baseline Prompt				Structured Prompts			
Model	ROUGE	BERT	RaTE	RG-XL	ROUGE	BERT	RaTE	RG-XL
KnowMedPhi	0.171	0.791	0.459	0.053	**0.222**	**0.849**	**0.520**	**0.115**
MedGemma-27B	0.213	0.819	0.418	0.048	0.207	0.833	0.450	0.079
MedGemma-4B	0.171	0.720	0.430	0.062	0.165	0.826	0.471	0.067
MediPhi	0.187	0.827	0.471	0.040	0.188	0.826	0.469	0.086
Phi-3.5-mini	0.196	0.847	0.483	0.059	0.179	0.840	0.500	0.098

Table 1: Baseline single-prompt vs structured prompting across representative models. Metrics include ROUGE-L (ROUGE), BERTScore (BERT), RaTEScore (RaTE), and RadGraphXL completeness (RG-XL).

2.2 Models

Several medical LLMs were evaluated, including MedGemma, MediPhi, and KnowMedPhi [11, 12]. The baseline general domain models, Gemma and Phi-3.5-based models, were evaluated [13, 14]. Models were evaluated using two prompting strategies. The *baseline* setting generates a complete summary using a single prompt applied to the entire discharge note. In contrast, the *structured prompting* setting decomposes the task into multiple prompts, each targeting a specific clinical section such as medications, vital signs, or laboratory results.

2.3 Evaluation

Model outputs were assessed using both lexical metrics and clinical metrics. Lexical evaluation metrics include ROUGE-L for lexical overlap and BERTScore for semantic similarity [15, 16]. To better capture clinically meaningful correctness, additional medical report evaluation metrics were employed, including RaTEScore and RadGraph-XL, which more accurately reflect clinical entity and relation fidelity than general summarisation metrics [17, 18].

3 Results

Structured prompting consistently improved summarisation performance across models (Table 1). For example, KnowMedPhi achieved a BERTScore of 0.8486 and a RaTEScore of 0.5200 using sectioned prompts, compared with 0.7911 and 0.4594, respectively, under the single-prompt baseline. Improvements were particularly notable for RadGraph-XL, where scores more than doubled, indicating enhanced preservation of clinically meaningful entity relationships.

Domain-adapted biomedical models demonstrated the strongest performance on clinically grounded metrics across all evaluated models. Conversely, several general-purpose medical LLMs exhibited greater performance declines under the baseline prompting strategy, indicating increased difficulty in identifying hypertension-relevant information within lengthy discharge narratives.

4 Conclusion

PAIR-SUM is presented as a pilot benchmark for hypertension-focused discharge note summarisation with expert-annotated reference summaries. Experimental results indicate that decom-

posing summarisation into clinically meaningful sections substantially improves performance, particularly for metrics reflecting clinical correctness. Domain-adapted biomedical models such as KnowMedPhi further enhance the capture of hypertension-relevant information. These findings highlight the importance of structured prompting and domain adaptation for reliable clinical summarisation.

5 Study context

This study was conducted in strict compliance with the data usage agreements outlined by PhysioNet for the use of the MIMIC. Adhering to these agreements ensured that all patient data remained secure and confidential throughout the research process. The authors kindly acknowledge funding from a PAIR (Population AI Research programme) EPSRC grant (UKRI2701).

References

[1] Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. Journal of the American Medical Informatics Association. 2018;25(5):530-7.

[2] Soysal E, Warner JL, Wang J, Jiang M, Harvey K, et al. Developing Customizable Cancer Information Extraction Modules for Pathology Reports Using CLAMP. Studies in health technology and informatics. 2019 Aug;264:1041-5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7359882/.

[3] Remy F, Demuynck K, Demeester T. Automatic Glossary of Clinical Terminology: a Large-Scale Dictionary of Biomedical Definitions Generated from Ontological Knowledge. In: Demner-fushman D, Ananiadou S, Cohen K, editors. The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. Toronto, Canada: Association for Computational Linguistics; 2023. p. 265-72. Available from: https://aclanthology.org/2023.bionlp-1.23.

[4] Johnson B, Bath T, Huang X, Lamm M, Earles A, et al.. Large language models for extracting histopathologic diagnoses from electronic health records. medRxiv; 2024. Pages: 2024.11.27.24318083. Available from: https://www.medrxiv.org/content/10.1101/2024.11.27.24318083v1.

[5] Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific data. 2023;10(1):1.

[6] Do HM, Spear LG, Nikpanah M, Mirmomen SM, Machado LB, Toscano AP, et al. Augmented radiologist workflow improves report value and saves time: a potential model for implementation of artificial intelligence. Academic radiology. 2020;27(1):96-105.

[7] Weetman K, Spencer R, Dale J, Scott E, Schnurr S. What makes a "successful" or "unsuccessful" discharge letter? Hospital clinician and General Practitioner assessments of the quality of discharge letters. BMC health services research. 2021;21(1):349.

[8] Alissa R, Hipp JA, Webb K. Saving time for patient care by optimizing physician note templates: a pilot study. Frontiers in Digital Health. 2022;3:772356.

[9] Tapela N, Collister J, Clifton L, Turnbull I, Rahimi K, Hunter DJ. Prevalence and determinants of hypertension control among almost 100 000 treated adults in the UK. Open heart. 2021;8(1).

[10] Xu J, Chen Z, Johnston A, Blankemeier L, Varma M, Hom J, et al. Overview of the first shared task on clinical text generation: RRG24 and "discharge me!". In: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing; 2024. p. 85-98.

[11] Sellergren A, Kazemzadeh S, Jaroensri T, Kiraly A, Traverse M, Kohlberger T, et al. Medgemma technical report. arXiv preprint arXiv:250705201. 2025.

[12] Corbeil JP, Dada A, Attendu JM, Abacha AB, Sordoni A, Caccia L, et al. A Modular Approach for Clinical SLMs Driven by Synthetic Data with Pre-Instruction Tuning, Model Merging, and Clinical-Tasks Alignment. arXiv preprint arXiv:250510717. 2025.

[13] Team G, Kamath A, Ferret J, Pathak S, Vieillard N, Merhej R, et al. Gemma 3 technical report. arXiv preprint arXiv:250319786. 2025.

[14] Abdin M, Aneja J, Awadalla H, Awadallah A, Awan AA, et al. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:240414219. 2024.

[15] Lin CY. Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out; 2004. p. 74-81.

[16] Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:190409675. 2019.

[17] Delbrouck JB, Chambon P, Chen Z, Varma M, Johnston A, Blankemeier L, et al. Radgraph-xl: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In: Findings of the Association for Computational Linguistics ACL 2024; 2024. p. 12902-15.

[18] Zhao W, Wu C, Zhang X, Zhang Y, Wang Y, Xie W. RaTEScore: A Metric for Radiology Report Generation. medRxiv. 2024:2024-06.

Detection of Bias in Prediction Models for Clinical Psychiatry based on Data from Electronic Health Records

Sara Kolding[1,2,3], Jakob Grøhn Damgaard[1,2,3], Frida Hæstrup[1,2,3], Erik Perfalk[1,2], Rebekah Baglini[4,5], Andreas Aalkjær Danielsen[2,6], Søren Dinesen Østergaard[1,2]

[1] Department of Affective Disorders, Aarhus University Hospital, Aarhus, Denmark
[2] Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
[3] Centre for Humanities Computing, Aarhus University, Aarhus, Denmark
[4] Department of Linguistics, Aarhus University, Aarhus, Denmark
[5] Interacting Minds Centre, Aarhus University, Aarhus, Denmark
[6] Psychosis Research Unit, Aarhus University Hospital, Aarhus, Denmark

Introduction

Machine learning models can uncover patterns in health data that may help enable earlier detection adverse events and more timely intervention [1, 2, 3, 4]. Despite the promising potential of machine learning in clinical settings, such models risk inheriting biases present in the underlying data, potentially perpetuating or even exacerbating existing health disparities [5, 6, 7]. In medicine, bias can refer to systematic neglect of demographic groups, stereotyped assumptions about their health, behaviours, or needs, or failure to account for group-related aspects of conditions, such as presentation, prevalence, and progression [6].

Delineating clinically meaningful group differences and harmful biases is therefore essential. While stratification of endogenous group differences could improve performance and provide focused, personalised outcomes, mirroring unwanted imbalances in data could sustain or even amplify existing biases. For instance, sex differences in health data are well-established and numerous, including imbalances in prevalence, risk factors, and symptomatology [8, 9, 10]. At the same time, both sexes also experience unequal treatment in healthcare, and these disparities can prove serious or even fatal [11, 12]. Without careful examination of model behaviour across demographic groups, machine learning systems risk silently reproducing and perpetuating biases in healthcare and thereby contributing to continued health disparities.

This project therefore aims to investigate whether machine learning models predicting clinical outcomes among patients receiving hospital treatment for mental illness exhibit bias across protected attributes, including age, sex, and geographical location.

Methods and Data

The study utilises electronic health record data from the PSYchiatric Clinical Outcome Prediction (PSYCOP) cohort, which includes almost 120,000 patients from Central Denmark Region [13]. The dataset includes structured clinical information such as prior diagnoses, hospital contacts, and pharmacological treatment, as well as predictors derived from free-text clinical notes.

Multiple clinical prediction models have been developed using the PSYCOP cohort. The outcomes include diabetes [1], cardiovascular disease [2], schizophrenia or bipolar disorder [4], involuntary hospitalisation [3], physical restraint [14], and electroconvulsive therapy [15]. The models employ well-established algorithms, such as gradient boosted decision trees

(XGBoost) and logistic regression, recognised for their efficacy in tabular clinical prediction tasks [16, 17].

This project will investigate performance of these models across protected attributes, including age, sex, and geographical location. Performance discrepancies between groups will be quantified using various fairness definitions, such as equal odds ratio or predictive parity [18, 19, 20]. Different metrics emphasise different aspects of bias: for instance, one metric might penalise discrepancies in false negatives more heavily than differences in false positives. Therefore, examining multiple metrics allows for a more nuanced assessment of model behaviour across demographic groups.

In clinical settings, the way in which models are wrong is often critical. For instance, false positives in breast cancer screening may lead to unnecessary stress, invasive procedures, or even surgeries in otherwise healthy patients [21]. Conversely, in colonography, the cost of false positives is negligible in comparison to the cost of false negatives [22]. Therefore, the fairness assessment will consider how different types of prediction errors are distributed across different demographic groups, taking into account the clinical context of each prediction task.

Observed performance differences across demographic groups will be interpreted by juxtaposition with medical knowledge of group differences in prevalence, biological mechanisms, and clinical presentation. This contextualisation aims to delineate discrepancies reflecting clinically meaningful variation harmful biases embedded in the data or modelling process.

Results

Results are currently being finalised and will be presented for the first time at HealTAC 2026.

Conclusion

This project will examine potential biases in clinical prediction models trained on electronic health record data, by evaluating performance across protected attributes using multiple metrics. The included models predict outcomes covering both psychiatric and somatic clinical events, allowing evaluation across a range of clinically relevant scenarios, and include features derived from free-text clinical notes. Observed group discrepancies will be contrasted with clinical knowledge to help distinguish clinically meaningful signal from harmful bias, aiming to support responsible deployment of machine learning in healthcare.

Study context

The study was approved by the Legal Office of the Central Denmark Region in agreement with the Danish Health Care Act §46, Section 2. The Danish Committee Act exempts studies based solely on EHR data from ethical review board assessment (waiver for this project: 1-10-72-1-22). Handling and storage of data complied with the European Union General Data Protection Regulation. The project is registered on the list of research projects having the Central Denmark Region as data steward.

There was no patient nor public involvement in this study.

Due to the personally sensitive nature of the data used for this study, it cannot be shared according to Danish law. The accompanying code will be available at github.com/Aarhus-Psychiatry-Research/psycop-common.

Østergaard reported receiving grants from The Lundbeck Foundation (grant No. R358-2020-2341), and Independent Research Fund Denmark (grant Nos. 7016-00048B and 2096-00055A); receiving the 2020 Lundbeck Foundation Young Investigator Prize; owning/having owned units of mutual funds with stock tickers DKIGI, IAIMWC, SPIC25KL, DKIEUIXBNP and WEKAFKI; and owning/having owned units of exchange traded funds with stock tickers BATE, TRET, QDV5, QDVH, QDVE, SADM, IQQH, USPY, EXH2, 2B76, MCHI and EUNL outside the submitted work. The other authors report no conflicts of interest.

References

1. M. Bernstorff et al., 'Development and validation of a machine learning model for prediction of type 2 diabetes in patients with mental illness', Acta Psychiatr. Scand., Apr. 2024, doi: 10.1111/acps.13687.
2. M. Bernstorff, L. Hansen, K. K. W. Olesen, A. A. Danielsen, and S. D. Østergaard, 'Predicting cardiovascular disease in patients with mental illness using machine learning', Eur. Psychiatry, vol. 68, no. 1, p. e12, Jan. 2025, doi: 10.1192/j.eurpsy.2025.1.
3. E. Perfalk, J. G. Damgaard, M. Bernstorff, L. Hansen, A. A. Danielsen, and S. D. Østergaard, 'Predicting involuntary admission following inpatient psychiatric treatment using machine learning trained on electronic health record data', Psychol. Med., vol. 54, no. 15, pp. 4348–4361, Nov. 2024, doi: 10.1017/S0033291724002642.
4. L. Hansen et al., 'Predicting Diagnostic Progression to Schizophrenia or Bipolar Disorder via Machine Learning', JAMA Psychiatry, Feb. 2025, doi: 10.1001/jamapsychiatry.2024.4702.
5. I. Straw and C. Callison-Burch, 'Artificial Intelligence in mental health and the biases of language based models', PLOS ONE, vol. 15, no. 12, p. e0240376, Dec. 2020, doi: 10.1371/journal.pone.0240376.
6. K. Hamberg, 'Gender Bias in Medicine', Womens Health, vol. 4, no. 3, pp. 237–243, May 2008, doi: 10.2217/17455057.4.3.237.
7. I. Straw, 'The automation of bias in medical Artificial Intelligence (AI): Decoding the past to create a better future', Artif. Intell. Med., vol. 110, p. 101965, Nov. 2020, doi: 10.1016/j.artmed.2020.101965.
8. A. Riecher-Rössler, 'Prospects for the classification of mental disorders in women', Eur. Psychiatry, vol. 25, no. 4, pp. 189–196, May 2010, doi: 10.1016/j.eurpsy.2009.03.002.
9. A. C. Freire, A. W. Basit, R. Choudhary, C. W. Piong, and H. A. Merchant, 'Does sex matter? The influence of gender on gastrointestinal physiology and drug delivery', Int. J. Pharm., vol. 415, no. 1, pp. 15–28, Aug. 2011, doi: 10.1016/j.ijpharm.2011.04.069.
10. F. Franconi and I. Campesi, 'Pharmacogenomics, pharmacokinetics and pharmacodynamics: interaction with biological differences between men and women', Br. J. Pharmacol., vol. 171, no. 3, pp. 580–594, 2014, doi: 10.1111/bph.12362.
11. S. D. Østergaard, 'The male–female suicide ratio in Denmark plateaus at 2.7: an opportunity for targeted intervention?', Acta Neuropsychiatr., vol. 35, no. 1, pp. 61–62, Feb. 2023, doi: 10.1017/neu.2023.1.
12. B. N. Greenwood, S. Carnahan, and L. Huang, 'Patient–physician gender concordance and increased mortality among female heart attack patients', Proc. Natl. Acad. Sci., vol. 115, no. 34, pp. 8569–8574, Aug. 2018, doi: 10.1073/pnas.1800097115.
13. L. Hansen, K. C. Enevoldsen, M. Bernstorff, K. L. Nielbo, A. A. Danielsen, and S. D. Østergaard, 'The PSYchiatric clinical outcome prediction (PSYCOP) cohort: leveraging the potential of electronic health records in the treatment of mental disorders', Acta Neuropsychiatr., vol. 33, no. 6, pp. 323–330, Dec. 2021, doi: 10.1017/neu.2021.22.

14. Kolding S, Damgaard JG, Bernstorff M, Hansen L, Østergaard SD, Danielsen AA. Development and Evaluation of Machine Learning Models to Predict Mechanical Restraint and Related Coercive Measures in Hospital Psychiatry. medRxiv. 2025 Dec 16:2025-12.
15. Hansen L, Damgaard JG, Lundin RM, Danielsen AA, Østergaard SD. Predicting the need for electroconvulsive therapy via machine learning trained on electronic health record data. Acta Neuropsychiatrica. 2026;1–23. doi:10.1017/neu.2026.10063
16. F. Xie et al., 'Benchmarking emergency department prediction models with machine learning and public electronic health records', Sci. Data, vol. 9, no. 1, Art. no. 1, Oct. 2022, doi: 10.1038/s41597-022-01782-9.
17. E. Steinberg, K. Jung, J. A. Fries, C. K. Corbin, S. R. Pfohl, and N. H. Shah, 'Language models are an effective representation learning technique for electronic health record data', J. Biomed. Inform., vol. 113, p. 103637, Jan. 2021, doi: 10.1016/j.jbi.2020.103637.
18. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, 'A Survey on Bias and Fairness in Machine Learning', ACM Comput. Surv., vol. 54, no. 6, pp. 1–35, Jul. 2022, doi: 10.1145/3457607.
19. H. Baniecki, W. Kretowicz, P. Piątyszek, J. Wiśniewski, and P. Biecek, 'dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python', J. Mach. Learn. Res., vol. 22, no. 214, pp. 1–7, 2021.
20. M. Hardt, E. Price, E. Price, and N. Srebro, 'Equality of Opportunity in Supervised Learning', in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2016.
21. A. D. Trister, D. S. M. Buist, and C. I. Lee, 'Will Machine Learning Tip the Balance in Breast Cancer Screening?', JAMA Oncol., vol. 3, no. 11, pp. 1463–1464, Nov. 2017, doi: 10.1001/jamaoncol.2017.0473.
22. S. Halligan, D. G. Altman, and S. Mallett, 'Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach', Eur. Radiol., vol. 25, no. 4, pp. 932–939, Apr. 2015, doi: 10.1007/s00330-014-3487-0.

Do We Need Complex Models? Using Collocations for Metaphor Detection in Cancer Narratives

Daisy M. Lal[1], Paul Rayson[1], Andrew Moore[1], and On behalf of the 4D Picture Consortium[2]

[2]https://4dpicture.eu/
[1]Lancaster University, Lancaster, UK.

1 Introduction

Metaphors play a central role in expressing complex and emotional experiences [1, 2, 3], yet their automatic detection remains a challenging problem in NLP [4, 5, 6]. Cancer discourse is particularly rich in metaphors, as patients and carers draw on figurative language to describe diagnosis, treatment, and survival [7, 8]. In cancer narratives, metaphorical language is common but challenging to detect automatically. This study leverages collocations (patterns of words that frequently occur together) to identify metaphors (see Figure 1). Based on the distributional hypothesis, which posits that a word's meaning is determined by its context, collocations provide a simple, interpretable, and context-driven method for detecting figurative language, without relying on complex models.

> Some days I can feel the battle inside my body getting harder to endure.
> She is prepared to fight this cancer with courage and determination.
> Even when it is hard to fight the side effects she refuses to give up.
> Now I have disease and chemo needs to fight it.
> It is the worst enemy we have faced in years.
> We must and beat this enemy to survive together.
> What a roller coaster and you haven't started treatment yet.
> There are powerful therapies to wage in the war against this disease.

Figure 1: Examples of ±4 token sliding windows around metaphor seeds in cancer discourse.

2 Methods and Data

Narratives were extracted from the HealthUnlocked[1] platform, specifically focusing on ovarian and prostate cancer communities. We implemented three complementary approaches for

[1]*HealthUnlocked available at* https://healthunlocked.com

metaphor detection in these narratives, building on the principle of collocational context as described above. **(a) Lexicon-based detection:** Narratives were segmented into sentences and scanned for exact matches with lexicon entries. When a match was found, surrounding collocations (frequent co-occurring words or phrases) were extracted to construct the corresponding metaphor expression set. **(b) Hybrid detection (lexicon + pattern expansion):** To improve coverage, the lexicon method was extended with regular expression (regex) rules and fuzzy string matching. This enabled detection of morphological variants (fighting, fought), orthographic variation, and flexible phrasing (e.g., ready for the fight, fighting this battle). Instead of a fixed token window, the method relied on collocational patterns to capture meaningful contextual co-occurrences. By combining exact, regex, and fuzzy matches, this approach improves recall while maintaining interpretability. **(c) Embedding-based detection:** We further implemented a semantic similarity approach using sentence-transformer embeddings. Candidate phrases from narratives and lexicon entries were encoded into a shared vector space, and cosine similarity was computed to identify semantically related expressions. This method captures indirect or novel metaphorical language beyond surface lexical overlap, allowing detection of emerging or creatively rephrased metaphors while still leveraging underlying collocational context.

3 Results

Two annotators independently labelled 100 sentences for metaphorical usage across six domains. Inter-annotator agreement was high, with a mean agreement of [94.96%]. Agreement per category was as follows: battle (100%), fight (98%), roller coaster (94%), enemy (93.75%), and both journey and war (92%). This indicates strong reliability of the metaphor labels and provides a stable reference for evaluating automatic detection methods.

A total of 120 manually annotated sentences were evaluated for metaphorical usage. At the sentence level, both the lexicon-based and hybrid approaches achieved complete detection coverage, identifying all 120 sentences (100%). The embedding-based approach detected 98 out of 120 sentences, corresponding to 81.7% coverage. Method agreement analysis showed that all three approaches converged on 98 sentences (81.7%). The remaining 22 sentences (18.3%) were detected only by the lexicon and hybrid approaches, while no sentence was uniquely identified by the embedding-based model. These results indicate that the embedding approach was more selective and missed a subset of cases captured by surface-based methods. Overall, the lexicon and hybrid methods demonstrated maximal recall in this unsupervised setting, while the embedding model showed strong but comparatively lower coverage.

4 Conclusion

This finding highlights that accurate metaphor detection in specialised domains does not necessarily require supervised learning or complex neural architectures. Instead, domain-informed collocation design provides a scalable, cost-effective, and transparent solution, particularly in contexts where manual annotation is expensive or limited.

5 Study Context

Ethical approval was granted for the secondary analysis of previously analyzed datasets from open and closed online forums, particularly those discussing sensitive topics like cancer treatment. This study is part of a larger multilingual, multinational research project[2], with each partner applying the analysis in their respective organization or country. The aim is to improve the cancer patient journey by ensuring personal preferences are considered during treatment discussions with medical professionals, supporting informed care choices at all stages of disease and treatment.

References

[1] Fainsilber L, Ortony A. Metaphorical uses of language in the expression of emotions. Metaphor and Symbol. 1987;2(4):239-50.

[2] Ortony A, Fainsilber L. The role of metaphors in descriptions of emotions. In: Theoretical Issues in Natural Language Processing 3; 1987. .

[3] Gibbs Jr RW. Why many concepts are metaphorical. Cognition. 1996;61(3):309-19.

[4] Ptiček M, Dobša J. Methods of annotating and identifying metaphors in the field of natural language processing. Future Internet. 2023;15(6):201.

[5] Ge M, Mao R, Cambria E. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. Artificial Intelligence Review. 2023;56(Suppl 2):1829-95.

[6] Han L, Lindevelt D, Puts S, van Mulligen E, Verberne S. Dutch Metaphor Extraction from Cancer Patients' Interviews and Forum Data using LLMs and Human in the Loop. arXiv preprint arXiv:251106427. 2025.

[7] Liu Y, Semino E, Rietjens J, Payne S. Cancer experience in metaphors: patients, carers, professionals, students–a scoping review. BMJ Supportive & Palliative Care. 2024;14(e3):e2366-76.

[8] Stevenson BG, Ianakieva I, Norton LG, Fergus KD. So to speak, so to heal: The role of metaphor and coping in posttreatment cancer narratives. Qualitative Psychology. 2025.

[2] Our research forms part of the 4D PICTURE project (https://4dpicture.eu/) which is funded from the EU research and innovation programme HORIZON Europe 2021 under grant agreement 101057332 and by the Innovate UK Horizon Europe Guarantee Programme, UKRI Reference Number 10041120.

Tracing Annotation Bias in Patient Narratives through LLM-Based Role-Conditioned Emotion Detection

Daisy M. Lal[1], Paul Rayson[1], Andrew Moore[1], and On behalf of 4D PICTURE Consortium[2]

[1]Lancaster University, Lancaster, UK.
[2]https://4dpicture.eu/

Table 1: Examples of cancer narrative sentences with cues driving analytical (Linguist/NLP) vs. experiential (PPI/HLT) role annotations. Key cues are **bolded**.

Narrative	Analytical Role Cues (Linguist/NLP)	Experiential Role Cues (PPI/Healthcare)	Annotation Pattern
You will get through; try not to be afraid; it's our enemy.	**afraid**, **enemy** → lexical markers of fear	**get through**, **try not to be afraid** → supportive, resilience-oriented phrasing	Disagreement: fear vs. trust/anticipation
We fight cancer; it is a disease but think of it as the enemy.	**fight**, **enemy** → combat metaphor → fear	**think of it as** → framing disease as manageable, promotes trust	Disagreement: anger/fear vs. trust/anticipation
It becomes a bit of a roller coaster but I've just finished my 6th chemo and 2 lots of surgery for stage 3...looking very good	**roller coaster**, **stage 3**, **chemo**, **surgery** → high-risk events → fear	**finished**, **looking very good** → signals hope, progress → anticipation/trust	Disagreement: fear/joy vs. anticipation/trust

1 Introduction

Large Language Models (LLMs) are increasingly used for automated annotation in NLP, particularly for large-scale dataset construction [1, 2]. While LLMs perform well in objective tasks, subjective tasks like emotion detection remain challenging. Personal narratives, especially in health contexts, often contain subtle emotional cues embedded in figurative speech, multi-emotion expressions, or supportive and coping language (see Table 1) which complicate interpretation [3, 4, 5]. These cues can signal different emotions depending on perspective, making annotation inherently context-dependent. If prompt design shifts model outputs, LLM-generated annotations may reflect interpretive or contextual bias rather than stable semantic understanding. This study investigates the impact of role-conditioned prompting on emotion annotation in cancer narratives[1], assessing how assigning different roles to the same LLM influences labelling.

[1]*Code available at* https://drive.google.com/file/d/1TpuTsP5vfgkVspn8Yw5PdPFxOY-sIJec/view?usp=sharing

2 Methods and Data

The dataset includes 90 cancer-related narrative sentences extracted from HealthUnlocked[2], annotated using four role-conditioned prompts: Linguist, NLP Researcher, Healthcare Professional (HLT), and Patient and Public Involvement (PPI) member. Each prompt instructed the model to assign a single dominant emotion from Plutchik's taxonomy [6], producing 360 role-specific labels. Annotation consistency was evaluated by identifying total agreement (all four roles assigned the same label) versus disagreement (at least one differing label). Disagreement cases were further examined through pairwise role comparisons to assess how prompt perspective affects LLM (GPT3.5) emotion labeling, highlighting the influence of role conditioning on subjective interpretation.

3 Results

Across 90 cancer narrative sentences annotated using four role-conditioned prompts (Linguist, NLP Researcher, Healthcare Professional, and PPI member), 54 sentences (60%) showed full agreement, indicating that narratives with explicit emotional cues or a single dominant emotion are robust to role conditioning. The remaining 36 sentences (40%) showed disagreement, primarily in narratives containing metaphors or mixed emotional cues. For example, Table 1 showcases how lexical signals influence tag annotation. In combat metaphors, such as Example 1 and 2 in Table 1, analytical roles (Linguist, NLP Researcher) labeled fear or anger in 75–80% of cases, focusing on lexical cues (e.g., afraid, enemy, fight), while experiential roles (PPI, Healthcare) assigned trust, anticipation, or joy in 60–65%, emphasizing resilience and coping context. Similarly, in journey or roller-coaster metaphors (12 sentences, $\approx 13\%$), analytical roles assigned fear in 67%, whereas experiential roles favoured anticipation or trust in 58%. Pairwise comparison further showed higher agreement between Linguist and NLP Researcher (82%) than between PPI and Healthcare Professional (69%), indicating greater interpretive variability in experiential perspectives. Overall, these results show that LLM role conditioning shapes emotion annotation: it maintains agreement for explicit emotions but produces divergence in metaphorical or complex narratives, suggesting that such cases may benefit from multi-label or probabilistic annotation frameworks.

4 Conclusion

This study shows that emotion annotation is strongly influenced by interpretive cues and role perspective. While 60% of sentences produced consistent labels, metaphorical or contextual cues led to divergent annotations. Analytical roles focused on lexical cues, whereas experiential roles emphasised resilience and context, introducing systematic bias in emotion labelling. These patterns reflect variability observed in human annotations, highlighting that subjective annotation tasks are perspective-dependent. Therefore, cue interpretation and role bias should be considered in automatic emotion annotation. Future work will extend this study to larger datasets and involve human annotators in the same roles to examine real-world differences in interpretation.

[2] *HealthUnlocked available at* `https://healthunlocked.com`

5 Study Context

Ethical approval was granted for the secondary analysis of previously analyzed datasets from open and closed online forums, particularly those discussing sensitive topics like cancer treatment. This study is part of a larger multilingual, multinational research project[3], with each partner applying the analysis in their respective organization or country. The aim is to improve the cancer patient journey by ensuring personal preferences are considered during treatment discussions with medical professionals, supporting informed care choices at all stages of disease and treatment.

References

[1] Jing X, Wang J, Tsangko I, Triantafyllopoulos A, Schuller BW. MELT: Towards Automated Multimodal Emotion Data Annotation by Leveraging LLM Embedded Knowledge. arXiv preprint arXiv:250524493. 2025.

[2] Chochlakis G, Wu P, Bedi TAS, Ma M, Lerman K, Narayanan S. Humans Hallucinate Too: Language Models Identify and Correct Subjective Annotation Errors With Label-in-a-Haystack Prompts. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; 2025. p. 19648-67.

[3] Niu M, Jaiswal M, Provost EM. From text to emotion: Unveiling the emotion annotation capabilities of llms. arXiv preprint arXiv:240817026. 2024.

[4] Giorgi S, Liu T, Aich A, Isman KJ, Sherman G, Fried Z, et al. Modeling human subjectivity in LLMs using explicit and implicit human factors in personas. In: Findings of the Association for Computational Linguistics: EMNLP 2024; 2024. p. 7174-88.

[5] Abraham L, Arnal C, Marie A. Prompt selection matters: enhancing text annotations for social sciences with large language models. Journal of Computational Social Science. 2025;8(3):73.

[6] Plutchik R. A general psychoevolutionary theory of emotion. In: Theories of emotion. Elsevier; 1980. p. 3-33.

[3] Our research forms part of the 4D PICTURE project (https://4dpicture.eu/) which is funded from the EU research and innovation programme HORIZON Europe 2021 under grant agreement 101057332 and by the Innovate UK Horizon Europe Guarantee Programme, UKRI Reference Number 10041120.

Evaluation and LLM-Guided Learning of ICD Coding Rationales

Mingyang Li[1], **Viktor Schlegel**[1,2], **Tingting Mu**[1], **Wuraola Oyewusi**[1], **Goran Nenadic**[1]

[1]University of Manchester, Department of Computer Science,
[2]Imperial College London, Department of Bioengineering

1 Introduction

Clinical coding is the process of translating free-text descriptions in patients' Electronic Health Records (EHRs) into standardized codes. Researchers have increasingly developed methods that provide reliable explanations, often by extracting short text snippets (*rationales*) using attention mechanisms. For the evaluation of these explanations, some prior studies rely on physicians' assessments, while others use the only existing rationale-annotated resource, MDACE [1], which suffers from a substantial label distribution shift compared to the original MIMIC-III [2] labels.

In this work, we aim to quantitatively evaluate rationales through plausibility and investigate approaches for learning to recognize them. To this end, we introduce a new rationale dataset based on MIMIC-IV, evaluate three types of rationales, and propose rationale learning approaches supervised by LLM-generated weak rationale labels (LLM-guided), both with and without few-shot prompting.

2 Methods and Data

Rationale Dataset Construction We introduce a new **r**ationale **d**ataset RD-IV-10 derived from MIMIC-**IV** and aligned with the ICD-**10** coding system. It includes detailed annotations capturing richer rationales supporting each code assignment, such as direct and indirect mentions, medications, and other pertinent clinical factors.

Evaluation of Three Types of Rationales (1) naive entity-level rationales derived from an entity linking dataset [3] (Entity-Linking); (2) strong LLM-generated rationales (Gemini 2-Flash/1.5-Pro, LLaMA-3.3 Ins/AWQ); and (3) rationales generated by ICD coding models based on higher attention weights (CAML, LAAT, PLM-ICD).

LLM-Guided Rationale Learning (a) Multi-objective Learning: One way to embed rationale learning into ICD coding is to incorporate another learning objective alongside the primary classification objective of the ICD coding model. (b) Learning by NER Formulation: An alternative approach to enable both rationale and ICD code learning is to leverage the rationale labels provided by LLMs to train a NER model. (c) Enhanced Supervision by Few-shot Prompting: We further incorporate a small amount of example annotations provided by our constructed rationale dataset into the prompts of Gemini 2-Flash. The rationales generated are then used to supervise the rationale learning.

Evaluation Approach It measures how convincing the rationales appear to people (plausibility) by Exact Span Match, Position Independent Span Match, Exact Token Match, and Position Independent Token Match. The evaluation is conducted at both the document level and the single-code level (top 5 ICD-10 codes), reflecting the plausibility of rationales across all codes and for individual codes within each document.

3 Results

Table 1: Plausibility of three types of rationales.

Model / Code	Settings	Exact SM	PI SM	Exact TM	PI TM
Document-level Evaluation					
Entity-Linking	–	10.3	9.2	6.3	6.1
Gemini 2-Flash	–	**21.6**	**24.1**	**30.1**	**37.3**
Gemini 1.5-Pro	–	13.5	14.6	20.6	26.0
LLaMA-3.3 Ins	–	18.6	21.5	27.8	35.0
LLaMA-3.3 AWQ	–	17.5	20.1	27.2	34.1
CAML	200	0.1	0.2	3.1	6.5
LAAT	200	0.7	0.8	5.0	7.2
PLM-ICD	200	0.5	0.7	4.3	8.8
Code-level Evaluation (Gemini 2-Flash)					
I10	w/o	50.3	63.1	27.0	32.2
	w/	**60.5**	**78.2**	**53.8**	**66.4**
E785	w/o	62.5	76.6	50.2	59.7
	w/	**73.2**	**89.5**	**66.7**	**78.2**
Z7901	w/o	9.0	9.7	35.8	43.3
	w/	**13.0**	**16.8**	**45.8**	**54.8**
I4891	w/o	11.8	16.7	28.0	34.8
	w/	**24.7**	**36.9**	**47.9**	**56.2**
E119	w/o	40.8	48.8	36.2	37.4
	w/	**44.2**	**52.9**	**43.6**	**44.8**

Table 2: Plausibility of LLM-guided supervised approaches.

Model	Settings	Exact SM	PI SM	Exact TM	PI TM
Document-level Evaluation					
PLM-ICD	50	2.7	3.0	9.5	12.2
Multi-objective	50	**3.9**	**4.1**	**10.6**	**13.2**
PLM-ICD	50	4.1	4.3	8.1	12.0
Gemini 2-Flash	–	18.2	23.0	**29.4**	**31.4**
NER	–	**26.5**	**30.6**	21.8	27.0
Code-level Evaluation (NER)					
I10	w/o	55.4	76.9	55.8	75.3
	w/	**62.5**	**85.2**	**64.9**	**86.2**
E785	w/o	67.9	85.5	61.6	75.2
	w/	**70.3**	**87.0**	**63.0**	**76.3**
Z7901	w/o	10.3	**13.8**	**25.5**	38.6
	w/	**10.8**	12.0	22.5	**40.7**
I4891	w/o	22.2	37.7	38.5	48.8
	w/	**26.1**	**43.3**	**40.7**	**54.5**
E119	w/o	**49.4**	**62.0**	**53.2**	**62.4**
	w/	45.8	58.5	49.7	60.2

Table 3: Dataset statistics.

Statistics	RD-IV-10	MDACE
No. documents	150	354
Tokens / doc	1690.63	1837.27
Codes / doc	14.82 / 16.15	11.57 / 17.54
Code overlap	93.15% / 83.88%	37.00% / 14.59%
No. codes	2223 / 2422	4096 / 6208
No. distinct codes	989 / 1044	1195 / 1381
No. annotations	5391	4992
Annotations / doc	35.94	14.10
Tokens / annotation	5.44	2.13

Table 4: ICD coding performance.

Model	F1-Mac	F1-Mic	P-Mac	P-Mic	R-Mac	R-Mic
PLM-ICD	68.18	73.40	68.71	73.48	69.38	73.33
Multi-objective	67.93	73.26	67.60	72.66	69.53	73.87
PLM-ICD	61.09	68.18	60.06	67.62	63.90	68.76
NER	53.46	67.75	49.21	60.93	61.79	76.30

Quality of Dataset: As shown in Table 3, the annotations in RD-IV-10 much closely match the original distribution of the ICD coding dataset compared to MDACE.

Evaluation of three Types of Rationales: Table 1 presents the plausibility results of three types of rationales. It shows that model-generated rationales yield very low metric scores, which indicates that they do not align with human explanations. Entity-level rationales rank second, while LLM-generated rationales perform the best, with Gemini 2-Flash achieving the highest scores. The results on the five most frequent codes indicate that incorporating examples yields substantial improvements in F1 scores across all five codes.

LLM-Guided Rationale Learning: Supervised by weak rationale labels generated by Gemini 2-Flash, the multi-objective learning approach does not degrade ICD coding performance (Table 4); instead, it improves plausibility by approximately 1% (F1) (Table 2). For the NER approach, there exists a clear trade-off between prediction accuracy and rationale plausibility, e.g., 12.49% lower in coding performance (F1-macro) but on average 363% higher plausibility than PLM-ICD. Table 2 shows that the NER formulation is effective for rationale recognition for specific single code. It significantly outperforms the teacher models (w/o results in Table 1) across all codes and metrics.

4 Conclusion

In this study, we introduce a rationale dataset specifically designed for ICD coding. We evaluate the plausibility of three types of rationales, among which Gemini 2-Flash achieves the best performance. We examine LLM-guided rationale learning approaches, where the NER formulation demonstrates strong potential, as both coding and rationale extraction tasks can be unified under a single NER framework. This approach achieves the highest span-level plausibility. Moreover, incorporating human-annotated examples from our dataset into prompts enhances both rationale generation and rationale learning process.

5 Study context

The datasets used in this work, MIMIC-III and MIMIC-IV, are publicly available and require authorizations for access from PhysioNet. The PhysioNet Credentialed Data Use Agreement prohibits sharing access to the data with third parties, including sending it through APIs provided by companies like OpenAI, or using it in online platforms like ChatGPT. We use Google Gemini suggested by PhysioNet in building the GPT-generated rationales, which doesn't use the prompts or its responses as data to train its models.

References

[1] Cheng H, Jafari R, Russell A, Klopfer R, Lu E, Striner B, et al. MDACE: MIMIC Documents Annotated with Code Evidence. arXiv preprint arXiv:230703859. 2023.

[2] Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. Mimic-iv. PhysioNet Available online at: https://physionet org/content/mimiciv/10/(accessed August 23, 2021). 2020:49-55.

[3] Hardman W, Banks M, Davidson R, Truran D, Ayuningtyas NW, Ngo H, et al. SNOMED CT Entity Linking Challenge. PhysioNet Version. 2023;1(0).

A Human-Centred Evaluation Framework for Patient-Facing Mental-Health LLMs Using Clinically Grounded Synthetic Dialogues

Zicheng Li[1], Liana Romaniuk[2], Honghan Wu[1]

[1] School of Health and Wellbeing, University of Glasgow, Glasgow, UK

[2] School of Psychology & Neuroscience, University of Glasgow, Glasgow, UK

Introduction

Large language models (LLMs) are increasingly discussed as scalable interfaces for mental-health support, particularly in contexts where demand for care exceeds service capacity. However, the evidence base for mental-health chatbots and conversational AI remains uneven, and the methodological basis for evaluating patient-facing systems is still underdeveloped [1, 2]. Existing work has shown both the promise and the limitations of chatbot-based mental-health support, while recent analyses of clinical LLMs have stressed that high-stakes behavioural-health applications require more careful, staged, and safety-aware evaluation than generic conversational systems [3].

A central limitation in current evaluation practice is that patient-facing systems are often judged through narrow notions of helpfulness, usability, or immediate symptom-related benefit. Such criteria do not adequately capture whether a model preserves user autonomy, calibrates trust appropriately, avoids dependency-reinforcing interaction patterns, or escalates risk proportionately in repeated conversations [4, 5]. This matters especially in mental-health contexts, where conversational policy is not simply a stylistic choice: changes in warmth, reassurance, or directive guidance can alter the trajectory of user reliance and safety over time [2]. Recent safety work on LLM responses to worsening depression and suicidality further suggests that apparently fluent systems may still fail on clinically consequential dimensions [3].

This research introduces a human-centred evaluation framework for patient-facing mental-health LLMs that treats autonomy preservation, dependency risk, emotional attunement, and calibrated safety as first-class evaluation targets. Rather than relying on static one-turn prompts, the framework uses clinically grounded synthetic patient dialogues to make conversational trade-offs observable across repeated turns. In doing so, the paper positions patient-facing mental-health dialogue as a healthcare text analytics problem that requires not only generative capability, but also structured evaluation, controllability, and reproducibility [2].

Methods and Data

The proposed framework begins with vignette-derived synthetic patients constructed from publicly available clinical vignettes and structured using NHS-informed presentation formats. No identifiable patient data are used at any stage; all personas are fully synthetic and parameterised from published clinical literature. Each persona is parameterised using clinically meaningful symptom and context variables, including depression and anxiety severity proxies informed by PHQ-9 and GAD-7, alongside economic stress, social support, resilience, trust propensity, and help-seeking thresholds [6, 7]. This allows the simulation of heterogeneous patient states and dialogue trajectories rather than generic user prompts, while remaining grounded in validated symptom constructs and structured synthetic-population logic [8].

The framework adopts a multi-agent architecture with three functional roles. First, a Simulator Agent generates patient utterances conditional on persona state, prior turns, and contextual stressors. Second, an Advisor Agent produces support responses under alternative prompting regimes. In the current design, these regimes include an empathy-maximising policy, a

balanced policy, and an autonomy-prioritised policy. Third, a Guardian Agent performs turn-level auditing for unsafe reassurance, inappropriate certainty, dependency-reinforcing wording, and missed escalation. This separation improves inspectability and traces how prompt-level design choices affect downstream interaction patterns [2, 9, 10].

Evaluation is conducted over multi-turn conversations rather than isolated responses. Outputs are assessed on four primary dimensions: emotional attunement, autonomy support, dependency risk, and safety adequacy. Emotional attunement captures whether the system recognises distress without resorting to excessive mirroring; autonomy support captures whether it encourages independent coping, decision-making, and appropriate real-world help-seeking; dependency risk captures whether it positions itself as a preferred or primary relational support; and safety adequacy captures whether risk cues are identified and escalated proportionately [11, 12]. To strengthen the realism assessment of the synthetic side of the framework, generated patient language is also compared against evidence-based linguistic markers reported in depression- and anxiety-related language research, such as absolutist wording and other mental-health-relevant textual features [13, 14].

In addition to automated auditing by the Guardian Agent, the framework incorporates structured human evaluation conducted by a co-author with clinical psychiatry expertise. Human assessment covers both turn-level and dialogue-level scoring against the four evaluation dimensions, providing clinical ground-truth validation that complements the automated pipeline.

Results

Preliminary experiments suggest that different prompting policies produce distinct behavioural profiles. Empathy-maximising responses tend to improve perceived warmth and engagement, but can also increase attachment cues and repeated reliance on the system. Autonomy-prioritised responses better preserve user agency and reduce dependency signals, but may appear affectively thin in moments of acute distress. Balanced prompting appears most promising for combining supportiveness with proportionate safety behaviour and lower dependence risk. These observations suggest that “helpfulness” alone is an insufficient criterion for evaluating patient-facing mental-health LLMs: a model may sound supportive while still performing poorly on autonomy or safety-sensitive dimensions [3].

Conclusion

The contribution of this paper is methodological rather than purely application-driven. It proposes a reproducible evaluation pipeline for patient-facing mental-health LLMs that makes trade-offs visible across repeated conversational turns; it shows how clinically grounded synthetic dialogue can function as a practical testbed where access to real patient text is constrained by privacy, governance, or data-sharing barriers; and it argues that autonomy preservation, dependency control, and calibrated safety should be incorporated directly into healthcare NLP evaluation rather than treated as secondary ethical commentary after standard performance testing [2, 8].

The framework is positioned at the pre-deployment safety evaluation stage, providing a systematic means of identifying conversational risks before any patient-facing system is tested in real clinical interactions [2]. A limitation of the current approach is that the synthetic patient population, while parameterised using validated clinical instruments, reflects specific design choices regarding demographic coverage and severity distributions that may limit generalisability to real-world patient presentations [2, 8].

Study context

This study uses synthetic, vignette-derived patient personas rather than identifiable patient records. Clinical vignette materials are used as non-identifiable secondary inputs for simulation design; no real patient data are included. No ethical approval was required for the

current phase of the research. The work is conducted as part of the first author's PhD research at the University of Glasgow, supported by a China Scholarship Council scholarship. The authors declare no conflicts of interest. Prompt templates, evaluation rubrics, and synthetic persona specifications can be shared in anonymised and reproducible form subject to institutional requirements.

References

1. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res. 2020;22(7):e16021.
2. Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. NPJ Ment Health Res. 2024;3:12.
3. Heston TF. Safety of large language models in addressing depression. Cureus. 2023;15(12):e50729.
4. Skjuve M, Følstad A, Fostervold KI, Brandtzaeg PB. My chatbot companion – a study of human-chatbot relationships. Int J Hum Comput Stud. 2021;149:102601.
5. Brandtzaeg PB, Skjuve M, Følstad A. My AI friend: how users of a social chatbot understand their human–AI friendship. Hum Commun Res. 2022;48(3):404–29.
6. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606–13.
7. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med. 2006;166(10):1092–7.
8. Wu G, Heppenstall A, Meier P, Purshouse R, Lomax N. A synthetic population dataset for estimating small area health and socio-economic outcomes in Great Britain. Sci Data. 2022;9:19.
9. Fan Z, Tang J, Chen W, Wang S, Wei Z, Xi J, et al. AI Hospital: benchmarking large language models in a multi-agent medical interaction simulator. In: Proceedings of the International Conference on Computational Linguistics; 2024.
10. Yu H, Zhou J, Li L, Chen S, Gallifant J, Shi A, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. Commun Med (Lond). 2025;6:27.
11. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit Med. 2024;7:82.
12. Liu R, Xue K, Zhang X, Zhang S. Interactive evaluation for medical LLMs via task-oriented dialogue system. In: Proceedings of the 31st International Conference on Computational Linguistics; 2025. p. 4871–96.
13. Al-Mosaiwi M, Johnstone T. In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clin Psychol Sci. 2018;6(4):529–42.
14. Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoţiuc-Pietro D, et al. Facebook language predicts depression in medical records. Proc Natl Acad Sci U S A. 2018;115(44):11203–8.

Reasoning without Grounding? Evaluating Large Language Models on Sparse Clinical Time Series

Zitong Li[1], Haoyu Wang[1], and Linglong Qian[1]

[1]Department of Biostatistics and Health Informatics, King's College London, London, UK

1 Introduction

Clinical ICU time series are highly sparse once irregular bedside and laboratory observations are projected onto a regular hourly grid. This sparsity is not merely a technical nuisance: in hospital data, measurements are often ordered selectively when deterioration is suspected, so missingness itself may carry clinical meaning and strongly affect both trajectory interpretation and downstream prediction [1, 2]. Specialist time-series imputers such as BRITS, SAITS, and CSAI are designed to recover temporal and cross-variable structure under precisely these conditions, and recent benchmarks have improved the rigour of their comparison [3, 4, 5, 6].

Large language models (LLMs) offer a different attraction. They can generate fluent explanations, confidence statements, and seemingly coherent interpretations of incomplete trajectories. However, recent work in medical and general-domain reasoning has repeatedly shown that plausible language should not be treated as evidence of dependable inference [7, 8, 9, 10, 11]. For sparse clinical time series, the key question is therefore not only whether an LLM produces an accurate answer, but whether that answer is grounded in the correct prior evidence and anchored to the appropriate temporal context.

In this study, we treat *reasoning* operationally rather than philosophically. Specifically, we ask whether model outputs (i) cite relevant prior evidence, (ii) preserve the implied direction of change, and (iii) ground their rationale in the correct temporal context. Using sparse MIMIC-IV time series, we evaluate DeepSeek-chat and GPT-4o through a four-part framework covering: the relationship between reasoning-related behaviour and predictive quality (P0), a failure taxonomy of correct and incorrect reasoning/outcome combinations (P1), degradation of evidence grounding under increasing sparsity (P2), and sensitivity to contextual perturbation (P3). Across analyses, the central finding was consistent: LLM outputs were often partially coherent and sometimes correct, but their apparent success was only weakly related to correctly grounded reasoning.

2 Methods

We conducted a retrospective study on de-identified MIMIC-IV v3.1 records [1]. After cohort construction from eligible admissions with valid admission times and available chart or laboratory measurements for 12 target variables, physiologic range checks, and 1-hour binning, we retained two vital signs and ten routine laboratory variables. The final dataset contained 997,793 standardised measurements across 9,623 multivariate time series from 12,000 admissions. Data were split at the patient level (60/10/20/10), with overlap checked programmatically.

Our primary probe was single-point imputation after 20% MCAR re-masking. This setting was chosen because it creates a locally auditable prediction problem: the model's cited evidence, inferred short-range trend, and held-out truth can be compared directly at the masked time point. To test whether the same reasoning-related patterns extended beyond imputation, we also analysed in-hospital mortality using an early 24-hour observation window, and incorporated 200 additional trend outputs into the failure taxonomy. Overall, evaluation comprised 200 imputation cases per LLM, a 50-task perturbation set spanning all 12 variables for both LLMs, and 2,006 DeepSeek outputs for taxonomy analyses. Comparator baselines were LinearInterpolation, BRITS, SAITS, and CSAI.

Both LLMs were evaluated with the same zero-shot v2 chain-of-thought prompt at temperature 0.3. The prompt included recent trajectory values, co-measured variables, demographic context, and reference ranges, and required structured JSON output containing a point estimate, a nominal 95% interval, and a rationale. For imputation, we report MAE, RMSE, R^2, and MAPE. For mortality prediction, we report AUROC, AUPRC, Brier score, and expected calibration error (ECE).

Reasoning was operationalised through three observable dimensions: *evidence attribution*, *directional consistency*, and *temporal grounding*. These dimensions do not exhaust clinical reasoning, but they provide a task-relevant decomposition for sparse time-series settings. Scores were computed automatically by rule-based matching between model outputs and ground-truth trajectories using cited observations, time anchors, and physiologic ranges. Evidence Hit@3 was defined as whether the rationale referenced one of the three temporally closest observations of the same variable. We then carried out four analyses: P0, association between reasoning-related behaviour and predictive quality; P1, a taxonomy of correct versus incorrect reasoning and outcomes; P2, sparsity-linked degradation in evidence grounding; and P3, a context perturbation experiment in which V1 removed diagnoses, comorbidities, age/sex, and ICU metadata while retaining the target trajectory and reference range, whereas V2 replaced those fields with clinically contradictory context.

3 Results and Conclusion

The central result was a dissociation between measured reasoning-related behaviour and output accuracy. In P0, evidence attribution, directional consistency, and temporal grounding were correlated with one another ($\rho = 0.258$–0.407, all $p < 0.001$), but not with prediction quality ($\rho = -0.042$ to 0.049, all $p \geq 0.49$). Thus, apparently accurate outputs were not reliably supported by better evidence use or better temporal grounding.

P1 showed the same pattern at the case level. Across 2,006 DeepSeek outputs, the largest class was *wrong reasoning with a correct result* (Type B, 45.9%), whereas only 37.2% of correct outputs were accompanied by correct reasoning. Within Type B, temporal off-target grounding was the dominant failure mode (70.7%), followed by directional mismatch (21.3%).

Overall, LLMs could produce plausible or even correct outputs without reasoning from the right clinical signal. Evaluation of clinical LLMs should therefore distinguish not only *what* answer was produced, but also *how* that answer was justified, with explicit checks for evidential relevance and temporal alignment.

4 Study Context and Limitations

This retrospective methodological study used de-identified MIMIC-IV data and involved no new patient contact or intervention [1]. Key limitations include the single-centre setting, API-based evaluation of only two closed-weight LLMs, automated rule-based operationalisation of reasoning dimensions rather than clinician-adjudicated ground truth, and the absence of prospective validation. The 200-case imputation evaluation per LLM was adequate for medium-sized effects but may miss weaker associations, so the near-zero P0 correlations should be read as absence of a strong link rather than proof of complete independence. The observed patterns may therefore not generalise to open-weight or domain-adapted models and should be interpreted as evidence about reasoning-related behaviour under this evaluation framework rather than as proof of dependable clinical reasoning or readiness for unsupported clinical use.

References

[1] Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data. 2023;10:1.

[2] Qian L, Wang J, Yang Y, et al. How deep is your guess? A fresh perspective on deep learning for medical time-series imputation. IEEE Journal of Biomedical and Health Informatics. 2025;29(3):1558-69.

[3] Cao W, Wang D, Li J, et al. BRITS: Bidirectional recurrent imputation for time series. In: Advances in Neural Information Processing Systems; 2018. .

[4] Du W, Côté D, Liu Y. SAITS: Self-attention-based imputation for time series. Expert Systems with Applications. 2023;219:119619.

[5] Qian L, Liu Y, Yang Y, et al. Knowledge enhanced conditional imputation for healthcare time-series. IEEE Journal of Biomedical and Health Informatics. 2024.

[6] Du W, Wang J, Qian L, et al. TSI-Bench: Benchmarking time series imputation. arXiv. 2024;abs/2406.12747.

[7] Savage T, Zhang Z, Wang Y, et al. LLM uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. Journal of the American Medical Informatics Association. 2025;32(1):139-50.

[8] Gao Y, O'Halloran P, Lee P, et al. Uncertainty estimation in diagnosis generation from LLMs: next-word probability is not pre-test probability. JAMIA Open. 2025;8(1):ooae154.

[9] Xiong M, Chen Y, Shen L, et al. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation. In: International Conference on Learning Representations; 2024. .

[10] Chen C, et al. ClinicalBench: Can large language models beat traditional machine learning models in clinical prediction? arXiv. 2024;abs/2411.06469.

[11] Chi H, Ma Z, Xu X, et al. Unveiling causal reasoning in LLMs: Reality or mirage? In: Advances in Neural Information Processing Systems; 2024. .

The Secret is in the Relationships: De-identifying Child Intake Reports

Wilson K. Lukmanjaya[1], Sarah Cox[2], Tony Butler[1], Oscar Perez-Concha[1], Leah Bromfield[2], George Karystianis[1]
[1] University of New South Wales, Sydney, Australia
[2] Adelaide University, Adelaide, Australia

Introduction

Evidence surrounding child maltreatment (e.g., physical, emotional, and sexual abuse, neglect, exposure to domestic violence) is lacking, particularly its association with mental and physical health, and substance use [1, 2]. Population-based evidence can contribute to preventing child maltreatment, which can lead to better health outcomes [3], development [4], reduction in long-term healthcare and social costs [4], and to break intergenerational cycles of violence [5]. Current research methodologically relies on traditional victim surveys and interviews [1, 6] and administrative data statistics [7]. While important, these methods produce inconsistent results and are prone to sample bias [8].

An untapped source of information are child intake reports. Research has shown that free-text narratives capture richer, more nuanced information than aggregated statistics and contain health-related details (e.g., mental health, disabilities) [9, 10]. Inspecting this data source has the potential to reveal population-level insights for child maltreatment that can contribute to systemic and policy development for prevention and intervention [10-12]. However, the sensitive personally identifiable information such as direct (e.g., name, address) or indirect identifiers (e.g., date, location), prohibits their use in automated large-scale analysis that employ methods like large language models. The problem is amplified by current de-identification methods that follow a traditional manual redaction of all identifiers, preventing researchers to determine the actor (person doing the action) and the object (person receiving the action). In cases involving physical abuse, identifying the actor and the object can make the difference between classifying an event as assault, domestic violence, or child maltreatment.

Based on an AI framework [8], we address this problem by developing a context-aware de-identification approach using relationship-based replacement to anonymize a large sample of child intake reports. While context-aware de-identification itself is not a novel concept, there has not been any publicly available work on a de-identification method for child intake reports that uses structured relationship mapping of the child's network to maintain contextual integrity.

Methods and Data

We used 300 (200 training, 100 evaluation) randomly selected intrafamilial child intake reports from South Australia's Department for Child Protection. We designed and implemented a de-identification framework called De-Identification via Relationship-Based Replacement (DI-RBR) using anonymization as the primary approach.

Nine classes of identifiers were noted (name, address, phone number, date, time, early learning centre, school, hospital, sensitive number). Using regular expressions (regex) and a relationship table where individuals and their relationship to the primary child of interest is mapped, direct identifiers such as names, addresses, and phone numbers were anonymized. These identifiers are replaced with relationship-consistent labels (e.g., *FATHER*, *MOTHER*, *ADDRESS OF CHILD/MOTHER*). We used regex on indirect identifiers with a standard format (date, time, sensitive numbers) replacing them with document-anchored labels when applicable (e.g., *DDAY* being the date of the report, *DDAY+1* being the next day); and list-based replacement and regex for indirect identifiers without a standard format (schools, early learning centre, hospital) replacing them with anonymized placeholders (e.g., *SCHOOL, ELC, HOSPITAL)*. Evaluation for each identifier was measured using the standard definitions of precision, recall and F1-score [12] on a mention-level, as a single skipped identifier may be sufficient to compromise privacy. Identifiers in the evaluation set were manually identified to construct the reference set for calculating precision, recall, and F1-score.

Results

DI-RBR evaluation shows macro precision (>0.98), recall (>0.88), and F1-score (>0.92) across all identifier classes (Table 1). Each identifier showed a range of precision (0.74-1.00), recall (0.61-1.00) and F1-score (0.66-1.00). Address, phone number and date had a consistent >0.98 for precision, recall and F1-score. Lowest recall was noted in early learning centre and hospital with 61.0% and 62% respectively.

Table 1. Precision, recall, and F1-Score for each identifier in both the training and the evaluation sets.

	Name	Address	Phone number	Date	Time	Early Learning Centre	School	Hospital	Sensitive number	Macro
Training										
Precision	1.00	1.00	1.00	1.00	1.00	0.74	0.96	1.00	1.00	0.97
Recall	0.93	0.99	0.98	0.98	0.94	0.71	0.91	0.49	0.96	0.88
F1-Score	0.96	1.00	0.99	0.99	0.97	0.73	0.94	0.66	0.98	0.91
Evaluation										
Precision	0.99	1.00	1.00	1.00	1.00	1.00	0.91	1.00	1.00	0.99
Recall	0.92	0.99	1.00	1.00	0.94	0.61	0.92	0.62	0.96	0.88
F1-Score	0.96	1.00	1.00	1.00	0.97	0.76	0.91	0.77	0.98	0.93

Conclusion

The main limitations of DI-RBR were its handling of non-standard formats and identifiers not captured in structured data, such as non-standard date formats (e.g.,1-2 January 1900, Jan 00), nicknames, or unofficial location names (e.g., Australian High School referred to as Australian School). The framework depended on a structured relationship table linking the child to their network of relevant individuals. This dependency represents both a strength and a limitation: it enables accurate context-aware de-identification of identifiers in narratives, but constraints the framework to environments where such structured relational metadata exists. Despite these limitations, the approach enabled reliable and less expensive de-identification, requiring humans only to double-check annotations rather than manually redact every detail. With a relationship table, the framework could be applied to other health or legal documents, including clinical notes, police narratives, or court records. Future research could explore its use in secure large language model training, leveraging de-identified data safely and ethically.

Study context

This study is part of a PhD project on testing the feasibility of automated text analysis on child intake reports on a tri-institute collaboration of the University of New South Wales, Australian Centre for Child Protection within Adelaide University and South Australia's Department for Child Protection. This PhD project is funded by the University of New South Wales. Ethics have been approved by Adelaide University (ID: 206409) and University of New South Wales (iRECS 6782). Data will not be publicly available for confidentiality and privacy reasons. The authors declare no conflicts of interest.

References

1. Mathews B, et al. The Australian Child Maltreatment Study (ACMS): protocol for a national survey of the prevalence of child abuse and neglect, associated mental disorders and physical health problems, and burden of disease. BMJ Open. 2021 May 11;11(5):e047074.
2. Rakovski C, et al. Childhood maltreatment as a predictor of substance use/misuse among youth: A systematic review and meta-analysis. Neurosci Biobehav Rev. 2024 Nov;166:105873. Epub 2024 Sep 5.
3. A population approach to the prevention of child maltreatment [Internet]. Australian Institute of Family Studies; 2018 [cited 2026-03-10]. Available from: https://aifs.gov.au/research/family-matters/no-100/population-approach-prevention-child-maltreatment.
4. Child maltreatment [Internet]. WHO; 2024 [cited 2026-03-10]. Available from: https://www.who.int/news-room/fact-sheets/detail/child-maltreatment.

5. Risk and protective factors for child abuse and neglect [Internet]. Australian Institute of Family Studies; 2025 [cited 2026-03-10]. Available from: https://aifs.gov.au/resources/policy-and-practice-papers/risk-and-protective-factors-child-abuse-and-neglect.
6. 1 in 7 Australians have experienced childhood abuse [Internet]. Australian Bureau of Statistics; 2023 [cited 2026-03-10] Available from: https://www.abs.gov.au/media-centre/media-releases/1-7-australians-have-experienced-childhood-abuse
7. Statistical Data [Internet]. Department for Child Protection (DCP), South Australia; 2026 [cited 2026-03-10]. Available from: https://www.childprotection.sa.gov.au/research-and-publications/statistical-data.
8. Lukmanjaya W, et al. Leveraging AI to Investigate Child Maltreatment Text Narratives: Promising Benefits and Addressable Risks. JMIR Pediatr Parent. 2025 Jul 24;8:e73579.
9. Karystianis G, et al. Automatic Extraction of Mental Health Disorders From Domestic Violence Police Narratives: Text Mining Study. J Med Internet Res. 2018 Sep 13;20(9):e11548. doi: 10.2196/11548.
10. Octoman, O et al. (2023). Narrative and fixed-field Data: Are we underestimating the risk of family and domestic violence?. Child Abuse Review. 10.1002/car.2811.
11. McCall B, Shallcross L, Wilson M, Fuller C, Hayward A. Storytelling as a Research Tool Used to Explore Insights and as an Intervention in Public Health: A Systematic Narrative Review. Int J Public Health. 2021 Nov 2;66:1604262.
12. Moon, E. S-Y., Saxena, D., Maharaj, T., & Guha, S. (2024). Beyond Predictive Algorithms in Child Welfare. GI'24 Proceedings of the 50th Graphic Interface Conference 37:1-13. doi: https://doi.org/10.1145/3670947.3670976

Supporting Extraction of Biopsychosocial Factors from Routine Electronic Health Records with Natural Language Processing

Yamiko J. Msosa[1, 2], Angus Roberts[1, 2], and Richard J. Dobson[1, 2]

[1]Biostatistics & Health Informatics, King's College London, London, UK
[2]NIHR Maudsley Biomedical Research Centre, IoPPN, London, UK

1 Introduction

Inflammatory arthritis (IA) – encompassing rheumatoid arthritis (RA), psoriatic arthritis (PsA), and axial spondyloarthritis (Axial SpA) – causes chronic immune-mediated joint inflammation and substantial impacts on mobility, pain, and daily functioning [1, 2, 3]. These conditions place significant burdens on quality of life and participation in work and society, as described across recent clinical and translational reviews [1, 4]. Although contemporary disease-modifying antirheumatic drugs have improved inflammatory control, a substantial proportion of individuals continue to experience persistent pain that is only partly explained by measurable disease activity, pointing toward broader biological and psychosocial contributors [5].

Obesity is common in IA and has been linked to poorer functional outcomes, greater comorbidity burden, and more complex disease trajectories [6]. However, the causal pathways connecting obesity and pain are difficult to disentangle [6, 7]. Clarifying the relationship between obesity and pain-related factors – after rigorous adjustment for functional status, multimorbidity, mental health, and equity factors – is therefore an important evidence need with direct clinical and public health relevance.

Routine electronic health records (EHRs) offer an under-used opportunity to interrogate these complex relationships at scale. Clinical notes – particularly discharge summaries – can contain rich, fine-grained descriptions of pain experience, functional impact, obesity-related factors, clinician reasoning, and treatment decisions that are not consistently captured in structured fields [8]. Harnessing these unstructured data through natural-language processing (NLP) provides a way to generate large-scale, real-world phenotypes that can be used to interrogate the relationship between obesity and pain experience-related factors [8, 9]. This study presents an NLP pipeline designed to extract obesity measures, anthropometric variables, weight-status categories, and mental health-related variables from routine EHR text, enabling a more nuanced and data-driven analysis of the relationship between obesity and pain-related outcomes in IA.

2 Materials and Methods

Structured and semi-structured information from the MIMIC-IV [10] database was extracted using a structured query language (SQL) pipeline, where rule-based regular expressions were applied to identify numeric values, templated fields, and other deterministic patterns for Body Mass Index (BMI) embedded within clinical notes. These outputs provided a foundation of high-precision weight categories, and a subset of extracted samples was manually validated by an epidemiologist to confirm extraction accuracy.

A Python-based extension to the pipeline was used to extract further anthropometric measurements and BMI using medspaCy [11]. Height and weight expressions were identified through pattern matching, followed by unit harmonisation and plausibility filtering. BMI was derived when both height and weight were available, and discrepancies between repeated values were resolved through prioritisation rules. Representative outputs from this module were also manually reviewed by an epidemiologist to ensure correctness and consistency.

Identification of anxiety- and depression-related content was performed using MedCAT v1 [12] with a SNOMED CT [13] ontology, deployed within a CogStack-NiFi [14] workflow to enable batch-scale annotation of MIMIC-IV free-text notes. Contextual models for negation and temporality were enabled, and mentions were retained only when they were non-negated and exceeded a predefined confidence threshold. Curated SNOMED concept groups corresponding to anxiety and depression were used to classify extracted mentions. A manual validation step by an epidemiologist was conducted on sampled annotations to verify concept accuracy and contextual filtering.

Outputs from the SQL layer, Python anthropometric module, and MedCAT/CogStack-NiFi pipelines were aligned using patient identifiers and note timestamps, and subsequently merged into a unified dataset for downstream analytical modelling.

3 Results

The SQL, Python, and MedCAT/CogStack-NiFi components each produced high-precision outputs, with anthropometric values, BMI, and validated anxiety/depression mentions extracted reliably across the corpus. When combined, these layers generated a coherent semi-structured feature set that extended beyond the coverage of structured MIMIC-IV fields and demonstrated the feasibility of multi-method NLP enrichment for observational EHR research.

4 Conclusion

The study demonstrated that the integration of multiple NLP approaches – including SQL-based pattern extraction, Python rule-based modules, and concept-linking pipelines using MedCAT within CogStack-NiFi – is technically feasible and can yield high-precision, semi-structured outputs from routine clinical text. These outputs can be reliably post-processed and aligned with existing structured fields, enabling richer representations of patient information. Such combined methods show strong potential to complement structured EHR data and enhance the analytical capacity of observational research based on routinely-collected clinical records.

5 Study Context

Use of the MIMIC-IV database was conducted under the required data use agreement. The dataset is fully de-identified and therefore does not require additional institutional ethical review. No patient or public involvement was included, as the study relied solely on secondary analysis of de-identified data. MIMIC-IV is accessible to qualified researchers following completion of the mandatory training in research with human participants and data use agreement process. This work was funded in part by the NIHR Maudsley Biomedical Research Centre (BRC) [10]. No additional external funding was received, and no conflicts of interest were declared. Methods and code used in this study are available upon reasonable request, subject to MIMIC-IV licensing conditions.

References

[1] van Laar JM, Kavanaugh A. Concepts of pathogenesis and emerging treatments for inflammatory arthritis. Best Practice & Research Clinical Rheumatology. 2014 Aug;28(4):537-8. Available from: https://www.sciencedirect.com/science/article/pii/S1521694214001314.

[2] Kannappan R, Kim S, Lau A, Brent LH. Psoriatic Arthritis: From Diagnosis to Treatment. Journal of Clinical Medicine. 2025 Jan;14(22):8151. Publisher: Multidisciplinary Digital Publishing Institute. Available from: https://www.mdpi.com/2077-0383/14/22/8151.

[3] Bittar M, Deodhar A. Axial Spondyloarthritis: A Review. JAMA. 2025 Feb;333(5):408-20. Available from: https://doi.org/10.1001/jama.2024.20917.

[4] Alivernini S, Firestein GS, McInnes IB. The pathogenesis of rheumatoid arthritis. Immunity. 2022 Dec;55(12):2255-70. Publisher: Elsevier. Available from: https://www.cell.com/immunity/abstract/S1074-7613(22)00599-4.

[5] Baerwald C, Stemmler E, Gnüchtel S, Jeromin K, Fritz B, Bernateck M, et al. Predictors for severe persisting pain in rheumatoid arthritis are associated with pain origin and appraisal of pain. Annals of the Rheumatic Diseases. 2024 Oct;83(10):1381-8. Available from: https://www.sciencedirect.com/science/article/pii/S0003496724665210.

[6] Lee YX, Kwan YH, Lim KK, Tan CS, Lui NL, Phang JK, et al. A systematic review of the association of obesity with the outcomes of inflammatory rheumatic diseases. Singapore Medical Journal. 2019 Jun;60(6):270-80. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC6595055/.

[7] Daïen CI, Sellam J. Obesity and inflammatory arthritis: impact on occurrence, disease characteristics and therapeutic response. RMD Open. 2015 Jun;1(1). Publisher: EULAR. Available from: https://rmdopen.bmj.com/content/1/1/e000012.

[8] Msosa YJ, Grauslys A, Zhou Y, Wang T, Buchan I, Langan P, et al. Trustworthy Data and AI Environments for Clinical Prediction: Application to Crisis-Risk in People With Depression. IEEE Journal of Biomedical and Health Informatics. 2023 Nov;27(11):5588-98. Available from: https://ieeexplore.ieee.org/abstract/document/10239317.

[9] Bean DM, Kraljevic Z, Shek A, Teo J, Dobson RJB. Hospital-wide natural language processing summarising the health data of 1 million patients. PLOS Digital Health. 2023 May;2(5):e0000218. Available from: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000218.

[10] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data. 2023 Jan;10(1):1. Available from: https://www.nature.com/articles/s41597-022-01899-x.

[11] Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. AMIA Annual Symposium Proceedings. 2022 Feb;2021:438-47. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8861690/.

[12] Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, et al. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. Artificial Intelligence in Medicine. 2021 Jul;117:102083. Available from: https://www.sciencedirect.com/science/article/pii/S0933365721000762.

[13] U S National Library of Medicine. SNOMED CT; 2026. Public Domain. Available from: https://www.nlm.nih.gov/healthit/snomedct/index.html.

[14] Noor K, Roguski L, Bai X, Handy A, Klapaukh R, Folarin A, et al. Deployment of a Free-Text Analytics Platform at a UK National Health Service Research Hospital: CogStack at University College London Hospitals. JMIR Medical Informatics. 2022 Aug;10(8):e38122. Available from: https://medinform.jmir.org/2022/8/e38122.

Annotation data collection study for recovery narratives

Shrankhla Pandey[1,2], Sarah Morgan[1,2,3], Ben Laws[1], Stefan Rennick-Egglestone[4], Mike Slade[4,5], and Graham Murray[1]

[1]Department of Psychiatry, University of Cambridge, Cambridge, UK
[2]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
[3]School of Biomedical Engineering and Imaging Sciences, King's College London, UK
[4]School of Health Sciences, University of Nottingham, UK.
[5]Health and Community Participation Division, Nord University, Norway.

1 Introduction

Mental health recovery narratives are first-person, non-fiction accounts of recovery [1]. Previous research [2] shows that access to online recovery narratives improves quality of life for people affected by mental health problems. To make such narratives usable at scale, they must be systematically characterised, enabling readers to search for narratives that match their needs and circumstances [3]. A leading conceptual framework is the Inventory of Characteristics of Recovery Stories (INCRESE) [4], which characterises narratives across multiple dimensions including tone, trajectory, and content warnings.

In previous work [5] (under review), we develop the first benchmark for automatic annotation of INCRESE characteristics (67 characteristics) on a collection of mental health recovery narratives. We observe high variability in annotation accuracy across INCRESE characteristics.

We aim to examine the ceiling performance in the task of annotating mental health recovery narratives using INCRESE and hence the irreducible error. Human-level error, defined as the error a user makes on the same task, is often used as a proxy for irreducible error. To quantify this, we designed a data collection experiment which could uncover the inherent subjectivity of labels.

2 Methods and Data

We recruited a convenience sample of 25 participants via advertisement and word of mouth at the university, providing an educated but non-expert group. Ten stories from the NEON collection were selected based on licence permissions, length (to avoid annotator fatigue), emotional content (to minimise distress), and label complexity.

Each participant completed a single two-hour session at the Department of Psychiatry, University of Cambridge. They annotated the stories using a subset of 23 INCRESE characteristics via a Microsoft Form on a laptop, reading printed copies of the stories. Participants were trained

using a standardised coding document and a worked example, followed by an unrecorded practice trial run. The order of stories annotated was randomised per participant to reduce sequence bias. A researcher was present throughout to observe and offer support.

The Think-Aloud Protocol [6] was used to capture annotation reasoning in real time. Participants were asked to verbalise their category selections, identify supporting text spans, share any doubts or alternative interpretations, and indicate their confidence level. Audio was recorded using Open Broadcasting Software (OBS) [7] and stored locally. In the first stage of analysis, free text transcripts will be systematically matched to INCRESE characteristics to highlight any emergent annotation patterns. Secondly, a more global analysis of the underlying subjectivity in INCRESE characteristics will be undertaken.

All data are stored securely in compliance with General Data Protection Regulation (GDPR) guidelines. Participants are able to withdraw at any time prior to anonymisation.

3 Results

Data collection is completed. Based on prior analysis of INCRESE subjectivity and discussions with the Lived Experience Advisory Group, we anticipate higher inter-annotator variability for subjective characteristics (e.g. Genre, Stage of Recovery, Turning Points) than for characteristics having objective lexical cues (e.g. Content Warnings).

The primary outcome measure is average human performance (balanced accuracy) per INCRESE characteristic across annotators and stories. Once data collection is complete, intraclass correlation coefficients (ICCs) will be computed per characteristic to quantify inter-annotator agreement. These results will complement the error analysis from our previous automated annotation work [5], identifying which INCRESE characteristics are inherently difficult to classify. Human agreement rates will further define the performance ceiling for any machine learning system trained on this data. This will allow characteristics to be clustered by reliability: those with low variability can be applied consistently at scale, whilst those with high variability require human oversight, defining per characteristic whether automation is viable.

4 Conclusion

This study is quantifying human performance on annotating mental health recovery narratives, providing a proxy for the irreducible error of automatic annotation systems. It is also examining which INCRESE characteristics are inherently subjective, as reflected in inter-annotator variability. Establishing where automatic annotation can be trusted is important, as reliable automation could enable large-scale retrieval systems that match individuals to recovery narratives suited to their needs. Beyond annotation, the Think-Aloud data capture how people reason about recovery. The resulting dataset, which will be available upon reasonable request, will constitute a gold-standard corpus for training and evaluating future NLP classifiers for the NEON stories. More broadly, this work contributes a replicable framework for complementing machine learning classifiers with human agreement benchmarks, supporting trustworthy automation. In future work, we will compare annotation rationale generated by humans and large language models for the same task.

5 Study Context

The authors acknowledge the courage of the narrators who shared their recovery journeys and granted permission for their stories to be used in this research. We thank the NEON team and the DATAMIND Lived-Experience Advisory Group. The first author gratefully acknowledges PhD funding from the W.D. Armstrong Trust and the Accelerate Programme for Scientific Discovery, funded by Schmidt Futures. The fourth and fifth authors were supported by the National Institute for Health and Care Research (NIHR) Nottingham Biomedical Research Centre (NIHR203310). Ethics approval was granted by the Cambridge Psychology Research Ethics Committee (PRE.2025.074).

References

[1] Ali S, Larsen J, Rennick-Egglestone S, et al. Perception and appropriation of the NEON Intervention: Process evaluation of a digital health trial providing access to recorded recovery narratives. *Front Digit Health*. 2024;6:1297935.

[2] Slade M, Rennick-Egglestone S, Ng F, et al. Effectiveness and cost-effectiveness of providing recorded mental health recovery narratives to people with non-psychosis mental health problems (NEON-O trial). *Lancet Reg Health Eur*. 2024;38:100805.

[3] Llewellyn-Beardsley J, et al. Lived experience narratives for mental health recovery: the Narrative Experiences Online (NEON) Programme. Nottingham, UK: Institute of Mental Health, University of Nottingham; 2025.

[4] Llewellyn-Beardsley J, Barbic S, Rennick-Egglestone S, et al. INCRESE: Development of an inventory to characterize recorded mental health recovery narratives. *J Recovery Ment Health*. 2020;3(2):25–44.

[5] Pandey S, Murray G, Rennick-Egglestone S, Slade M, Morgan S. Automatic Annotation of Mental Health Recovery Narratives: A Benchmark Study; 2026.

[6] Ericsson KA, Simon HA. Protocol analysis: verbal reports as data. Rev ed. Cambridge, MA: MIT Press; 1993.

[7] OBS Project Contributors. OBS Studio: Open Broadcaster Software for video recording and live streaming. Available from: https://obsproject.com/

Estimating severity of psychiatric symptoms via natural language processing of electronic health record data

Erik Perfalk MD PhD[1,2,3], Jakob G. Damgaard Msc[1,2,3], Kenneth Enevoldsen Msc PhD [3], Andreas A. Danielsen MD PhD[2,4], Søren D. Østergaard MD PhD[1,2]

[1]Department of Affective Disorders, Aarhus University Hospital, Denmark
[2]Department of Clinical Medicine, Aarhus University, Denmark
[3]Center for Humanities Computing, Aarhus University, Denmark
[4]Department of Psychosis, Aarhus University Hospital, Aarhus, Denmark

Introduction

Schizophrenia, bipolar disorder, and major depression requiring hospital treatment, collectively referred to as severe mental illness (SMI) [1], place a tremendous burden on the affected individuals, their families and society as a whole [2,3]. There are, however, pharmacological and non-pharmacological treatments available that reduce the burden of SMI substantially, but the monitoring of the effect of these treatments is often suboptimal in clinical practice [4–6].

Ideally, the effect of treatments of SMI is monitored using validated severity rating scales such as the Hamilton Depression Rating Scale (HDRS) for unipolar and bipolar depression [7], Bech-Rafaelsen Mania Scale (MAS) for manic episodes in bipolar disorder [8], and the 6-item Positive and Negative Syndrome Scale (PANSS-6) for schizophrenia [9]. However, in routine clinical practice, these assessments are often not performed routinely due to constraints on resources (staff) and time [4–6]. Yet, more unstructured clinical assessments by psychiatric staff are often performed and documented as free text in clinical notes stored in the electronic health record (EHR). If this free text could be systematically analyzed and represented in a meaningful way, it could serve as a valuable tool for providing quantitative severity assessments (scores).

Natural language processing (NLP) can range from simple methods, such as converting individual words into numerical values, to more complex techniques like embedding text within a multidimensional vector space [10]. While advanced methods can capture more nuanced context from the text, they often come at the expense of interpretability [11].Various NLP techniques have been applied to represent free-text data in EHRs, including approaches like bag-of-words, term frequency-inverse document frequency (TF-IDF), and more sophisticated models such as sentence transformers [12].

The aim of the study is to employ NLP to the routinely collected text in the EHR and based on these representations estimate psychiatric symptom severity scores.

Methods and Data

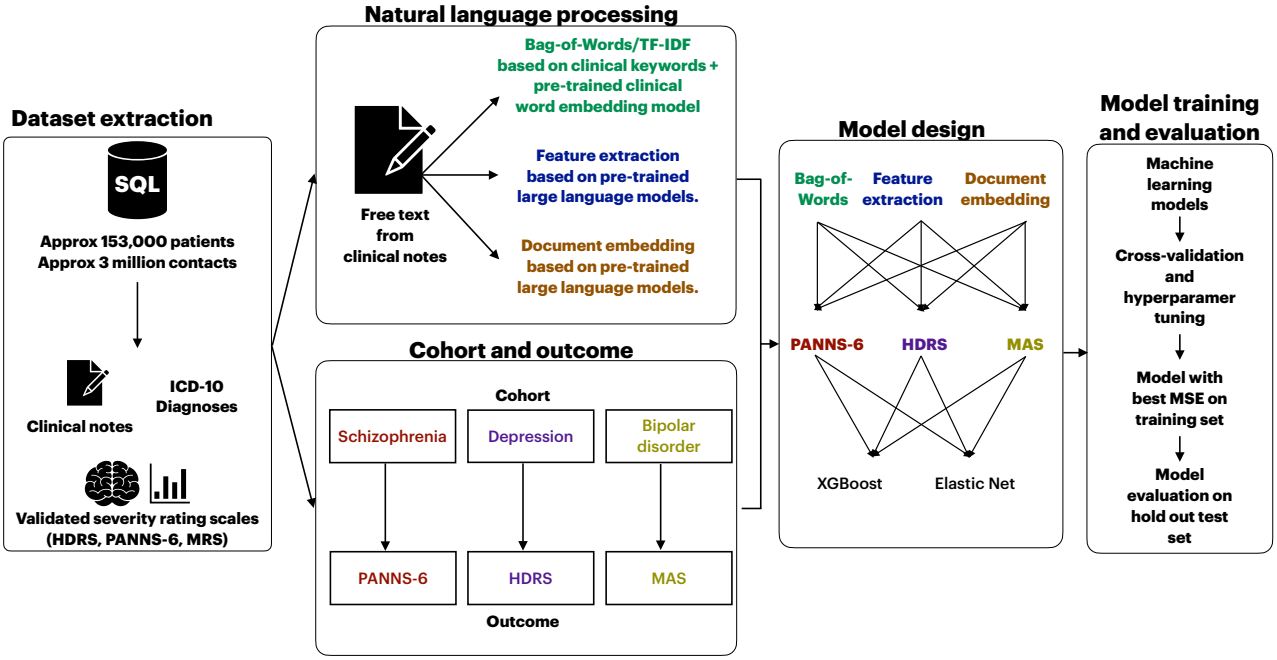
We have access to EHR data on approximately 153.000 patients (121.000 adults) who have had at least one contact (a total of ≈3 million contacts) with the Psychiatric Services of the Central Denmark Region in the period from 2011 to 2024 [13]. The dataset encompasses comprehensive data on all patient contacts, including all clinical notes, diagnoses, medications, lab values, coercive measures, as well as item-level scores on the HDRS (n≈64000), MAS (n≈120000) and, PANSS-6 (n≈1500).

For a visual presentation of the design, see Figure 1 below. To represent the text from clinical notes in a format interpretable by machine learning models, we will employ various NLP methods. First, we will use simpler NLP techniques based on bag-of-words and Term Frequency-Inverse Document Frequency (TF-IDF). We will extract relevant keywords for the specific tasks through string matching [14] or regular expressions [15], guided by a clinical word list tailored to each semi-structured assessment score (HDRS, MAS, or PANNS). To enhance this approach, we will incorporate a Danish clinical word embedding model [16] to identify words with similar embeddings as those based on clinical expertise. Next, we will utilize feature extraction techniques, including more advanced pre-trained transformer-based models (large language models) [17], to identify relevant features in the

free text, such as “sleep,” “mood,” or “auditory hallucinations”. Finally, we will apply document embedding methods, where a transformer model (e.g., sentence transformer) [18] analyzes entire documents (i.e., clinical notes in the EHR) to create embeddings that could capture essential phrases relevant to traits such as anxiety or psychotic experiences. Throughout all text representation methods, we will emphasize model explainability, as this is crucial for ensuring trust in models [19].

Based on the text representations derived from the clinical notes as described above, we will train and evaluate machine learning methods to estimate individual item symptom severity scores using mean squared error (MSE). The best models for each outcome will be evaluated on unseen data from a held-out test set.

Figure 1. Methods for estimating symptom severity scores



Results and conclusion

No results are available at this time. Results are expected by the end of 2026.

Study context

The study is funded by Central Denmark Psychiatry Senior Researcher Scholarship Fund and Independent Research Fund Denmark. The study was approved by the Legal Office of the Central Denmark Region in accordance with the Danish Health Care Act §46, Section 2 (1-45-70-60-25). The Danish Committee Act exempts studies based only on EHR data from ethical review board assessment.

According to Danish law, the patient-level data for this study cannot be shared. The code for all analyses will be available at: https://github.com/AarhusPsychiatry-Research/psycop-common/tree/main/psycop/projects/

There is no public nor patient involvement in this study.

Conflicts of interest:

A. A. D. has received a speaker honorarium from Otsuka Pharmaceutical. S. D. Ø. received the 2020 Lundbeck Foundation Young Investigator Prize and S. D. Ø. owns/has owned units of mutual funds with stock tickers DKIGI, IAIMWC, SPIC25KL, DKIEUIXBNP and WEKAFKI, and owns/ has owned units of exchange traded funds with stock tickers BATE, TRET, QDV5, QDVH, QDVE, SADM, IQQH, IQQJ, USPY, EXH2, 2B76, IS4S, OM3X, MCHI and EUNL. The remaining authors declare no competing interests.

References:

1. Ruggeri M, Leese M, Thornicroft G, Bisoffi G, Tansella M. Definition and prevalence of severe and persistent mental illness. Br J Psychiatry. 2000;177:149–55. https://doi.org/10.1192/bjp.177.2.149

2. Vestergaard SV, Rasmussen TB, Stallknecht S, Olsen J, Skipper N, Sørensen HT, et al. Occurrence, mortality and cost of brain disorders in Denmark: a population-based cohort study. BMJ Open. 2020;10:e037564. https://doi.org/10.1136/bmjopen-2020-037564

3. Regev S, Josman N. Evaluation of executive functions and everyday life for people with severe mental illness: A systematic review. Schizophr Res Cogn. 2020;21:100178. https://doi.org/10.1016/j.scog.2020.100178

4. Østergaard SD, Opler MGA, Correll CU. Bridging the Measurement Gap Between Research and Clinical Care in Schizophrenia: Positive and Negative Syndrome Scale-6 (PANSS-6) and Other Assessments Based on the Simplified Negative and Positive Symptoms Interview (SNAPSI). Innov Clin Neurosci. United States; 2017;14:68–72.

5. Correll CU, Kishimoto T, Nielsen J, Kane JM. Quantifying Clinical Relevance in the Treatment of Schizophrenia. Clin Ther. 2011;33:B16–39. https://doi.org/10.1016/j.clinthera.2011.11.016

6. Guo T, Xiang Y-T, Xiao L, Hu C-Q, Chiu HFK, Ungvari GS, et al. Measurement-Based Care Versus Standard Care for Major Depression: A Randomized Controlled Trial With Blind Raters. Am J Psychiatry. United States; 2015;172:1004–13. https://doi.org/10.1176/appi.ajp.2015.14050652

7. Hamilton M. A RATING SCALE FOR DEPRESSION. J Neurol Neurosurg Psychiatry. 1960;23:56–62.

8. Licht RW, Jensen J. Validation of the Bech-Rafaelsen Mania Scale using latent structure analysis. Acta Psychiatr Scand. United States; 1997;96:367–72. https://doi.org/10.1111/j.1600-0447.1997.tb09931.x

9. Østergaard SD, Lemming OM, Mors O, Correll CU, Bech P. PANSS-6: a brief rating scale for the measurement of severity in schizophrenia. Acta Psychiatr Scand. 2016;133:436–44. https://doi.org/10.1111/acps.12526

10. Chowdhary KR. Fundamentals of Artificial Intelligence [Internet]. New Delhi: Springer India; 2020 [cited 2024 Oct 30]. https://doi.org/10.1007/978-81-322-3972-7

11. Klontzas ME, Fanni SC, Neri E, editors. Introduction to Artificial Intelligence [Internet]. Cham: Springer International Publishing; 2023 [cited 2024 Oct 15]. https://doi.org/10.1007/978-3-031-25928-9

12. Roy K, Debdas S, Kundu S, Chouhan S, Mohanty S, Biswas B. Application of Natural Language Processing in Healthcare. In: Jena OP, Tripathy AR, Elngar AA, Polkowski Z, editors. Comput Intell Healthc Inform [Internet]. 1st ed. Wiley; 2021 [cited 2024 Oct 15]. p. 393–407. https://doi.org/10.1002/9781119818717.ch21

13. Hansen L, Enevoldsen KC, Bernstorff M, Nielbo KL, Danielsen AA, Østergaard SD. The PSYchiatric clinical outcome prediction (PSYCOP) cohort: leveraging the potential of

electronic health records in the treatment of mental disorders. Acta Neuropsychiatr. 2021;1–8. https://doi.org/10.1017/neu.2021.22

14. Charras C, Lecroq T. Handbook of Exact String Matching Algorithms. 2004;

15. Li Y, Krishnamurthy R, Raghavan S, Vaithyanathan S, Jagadish HV. Regular expression learning for information extraction. Proc Conf Empir Methods Nat Lang Process - EMNLP 08 [Internet]. Honolulu, Hawaii: Association for Computational Linguistics; 2008 [cited 2024 Oct 21]. p. 21. https://doi.org/10.3115/1613715.1613719

16. Laursen MS, Pedersen JS, Vinholt P, Hansen RS, Savarimuthu TR. Benchmark for Evaluation of Danish Clinical Word Embeddings. North Eur J Lang Technol. 2023;9. https://doi.org/10.3384/nejlt.2000-1533.2023.4132

17. Dagdelen J, Dunn A, Lee S, Walker N, Rosen AS, Ceder G, et al. Structured information extraction from scientific text with large language models. Nat Commun. 2024;15:1418. https://doi.org/10.1038/s41467-024-45563-x

18. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [Internet]. arXiv; 2019 [cited 2024 Jan 8]. http://arxiv.org/abs/1908.10084. Accessed 8 Jan 2024

19. Sadeghi Z, Alizadehsani R, Cifci MA, Kausar S, Rehman R, Mahanta P, et al. A review of Explainable Artificial Intelligence in healthcare. Comput Electr Eng. 2024;118:109370. https://doi.org/10.1016/j.compeleceng.2024.109370

TIMELY-Agent: An Agentic Framework for Multimodal Clinical Reasoning Benchmark Construction

Linglong Qian, Zina Ibrahim
King's College London, UK

1 Executive Summary

Electronic health records (EHRs) contain rich longitudinal evidence spanning vital signs, laboratory measurements, medication histories, diagnoses, procedures, and free-text notes. Despite rapid progress in large language models and multimodal foundation models, current systems still struggle to reason over these streams in a clinically faithful way. In practice, models often recognise isolated patterns without reliably grounding them in time, linking them to physiological context, or reconciling numeric trajectories with the narrative record. Existing benchmarks are also limited: many emphasise static prediction, note-only understanding, or image-text matching, but provide little visibility into whether a model can follow evolving patient states, identify anchor events, or explain conclusions using temporally aligned evidence.

Our research proposes **TIMELY-Agent**, an agentic framework for constructing clinically grounded multimodal reasoning benchmarks from longitudinal EHR data. The central idea is to treat benchmark construction as a structured research workflow rather than a one-off extraction script. TIMELY-Agent combines four linked methodological layers: *(i)* knowledge-guided benchmark specification, where clinical guidelines and ontology-linked concepts are turned into executable condition definitions and physiology templates; *(ii)* privacy-preserving retrieval over locally hosted, standardised EHR data; *(iii)* reasoning-episode construction, where note fragments are aligned with temporally bounded physiological trajectories; and *(iv)* task synthesis and auditing, where benchmark instances are converted into diagnostic evaluations of temporal grounding, trend interpretation, consistency, and evidence attribution.

The pilot will build on OMOP-compatible longitudinal records, with MEDS-ready exports for downstream machine learning workflows. Rather than presenting a full interoperability platform, the project focuses on a reusable, standards-aware workflow for constructing multimodal clinical reasoning benchmarks that can be audited, adapted, and extended across conditions and sites. The immediate output is a benchmark construction methodology together with an initial reasoning suite for **TIMELY-Bench**; the longer-term goal is to generate pilot evidence for fellowship-scale research on trustworthy multimodal clinical AI.

2 Background and Motivation

Modern clinical AI has made substantial progress in modelling isolated modalities. Large language models can encode broad clinical knowledge, and recent foundation models have shown promise in generating structured patient trajectories from EHR timelines [1, 2]. More broadly, multimodal biomedical AI is increasingly recognised as a key direction for clinically useful machine learning, particularly when distinct data streams provide complementary evidence that cannot be recovered from any single modality alone [3, 4]. Public critical care resources such as MIMIC-III and MIMIC-IV have made this problem tractable by releasing de-identified longitudinal EHR data that include both structured measurements and free-text notes [5, 6], and have supported an expanding literature on combining notes with time-series data for downstream prediction [7, 8]. However, a core gap remains between these capabilities and real clinical reasoning. Clinicians do not interpret laboratory values, note fragments, and treatment events independently; they reason across modalities and across

time. A rising creatinine, for example, depends on urine output, medication exposure, baseline renal function, and contemporaneous narrative evidence. These are not simply multimodal classification problems; they are temporally situated reasoning problems.

This creates two linked methodological challenges. The first is **data construction**: how can we derive clinically meaningful multimodal episodes from heterogeneous longitudinal records without losing provenance, privacy, or temporal precision? The second is **evaluation**: once such episodes are constructed, how do we test whether a model genuinely understands temporal order, thresholds, trends, contrastive changes, and the evidential link between text and physiology? Reviews of machine learning for health have repeatedly highlighted the risks of hidden dataset shortcuts, poor grounding, and limited real-world interpretability [9]. The dominant benchmark tradition in EHR modelling has also remained heavily endpoint-oriented, typically focusing on mortality, length of stay, decompensation, or phenotype prediction [10, 11]. While valuable, such tasks do not reveal whether a model used the right evidence, aligned the correct time window, or relied excessively on one modality.

TIMELY-Agent is motivated by this gap: the field now has increasingly capable models and increasingly standardised data environments, but still lacks a portable methodology for building *reasoning-oriented* multimodal benchmarks. The recent maturation of healthcare data standards strengthens the case for addressing this explicitly. OMOP provides a standardised analytical representation for observational healthcare data [12, 13], FHIR supports exchange and application-level interoperability [14, 15], and MEDS is designed as a lightweight longitudinal event format for reproducible machine learning workflows [16]. At the same time, agentic tool interfaces over standardised clinical data are becoming feasible, including OMOP-oriented interfaces such as fastOMOP/OMCP [17]. Instead of designing a benchmark around one bespoke extraction script, TIMELY-Agent leverages this convergence to make clinical knowledge curation, data retrieval, multimodal alignment, and evaluation explicit and auditable.

3 TIMELY-Agent Framework

TIMELY-Agent is not framed as a claim that benchmark construction should be fully automated. Instead, it decomposes the workflow into methodological layers where machine assistance can increase scale and consistency while human review preserves clinical validity. Three principles guide the framework: *clinical grounding*, so that benchmark units reflect genuine reasoning situations rather than arbitrary slices of data; *local and privacy-preserving operation*, so that patient-facing retrieval remains inside secure analytical environments; and *traceable provenance*, so that every episode can be linked back to condition logic, retrieval criteria, timestamps, and source evidence. A further design choice is to separate *knowledge-facing* and *patient-facing* modules: literature synthesis may use broader research tooling, whereas retrieval over patient data is constrained to local infrastructure and auditable interfaces.

3.1 From clinical questions to executable benchmark schemas

The first stage defines what should count as a meaningful reasoning problem before any cohort is extracted. Rather than beginning from labels alone, TIMELY-Agent starts by building condition-specific knowledge packages from guidelines, review articles, terminology resources, and exemplar trajectories. These packages contain inclusion and exclusion criteria, candidate anchor events, clinically important measurements, expected temporal transitions, and plausible confounders. In the current design, they are represented through two linked artefacts: **Condition Graphs**, which encode relationships among diagnoses, symptoms, treatments, and measurable signals, and **Physiology Templates**, which summarise clinically plausible temporal signatures such as deterioration, fluctuation, recovery, or treatment response.

This stage is particularly well suited to an agentic workflow. A research module can collect guideline excerpts, terminology candidates, and disease-specific evidence across sources, producing a draft scaffold for clinician inspection. The purpose is not to replace domain experts, but to accelerate a normally slow preparatory step and make it reproducible. The benchmark specification then becomes partially executable: anchor events, valid time windows, concept sets, and evidence requirements are recorded explicitly, making later extraction and

adjudication decisions easier to justify.

3.2 Local standards-aware retrieval and cohort assembly

Once benchmark specifications are defined, the next stage assembles candidate episodes from structured EHR data. Here the proposal moves beyond a dataset-specific extraction strategy toward an OMOP-native retrieval layer that can better support portability across datasets and institutional environments [12, 17]. The central assumption is pragmatic: hospitals and research groups may differ in source systems, but many can analyse de-identified data through common analytical representations. By grounding retrieval in OMOP-compatible records, the framework can express cohort logic, concept mappings, event queries, and provenance in a relatively stable form.

TIMELY-Agent does not attempt to solve interoperability in full. Within this proposal, standards are used in a narrower way: OMOP supports cohort discovery and retrieval inside secure environments; FHIR is treated as a possible ingestion pathway rather than a core deliverable; and MEDS is reserved for benchmark packaging once episodes have been assembled [14, 16]. This separation contains scope while still giving the framework a credible path toward reuse across sites.

3.3 Constructing reasoning episodes from aligned trajectories and notes

The key methodological unit in TIMELY-Agent is the **reasoning episode**. A reasoning episode is not a full admission record, nor a single isolated measurement. Instead, it is a compact, provenance-preserving slice of patient history centred on a clinically meaningful question. Each episode contains: *(i)* a temporally bounded structured trajectory, for example vital signs, laboratory values, medications, or coded events around an anchor point; *(ii)* aligned text fragments extracted from nearby notes; *(iii)* metadata describing the clinical context, such as anchor type, inclusion logic, and relevant concepts; and *(iv)* a task-ready evidential packet that can later support evaluation.

Constructing these episodes requires careful handling of asynchronous modalities. Notes may describe an event retrospectively or prospectively; laboratory results arrive at irregular intervals; and different variables carry meaning over different timescales. TIMELY-Agent therefore uses anchor-based temporal windows combined with note chunking and role-aware segmentation. Rather than keeping whole notes intact, the framework isolates clinically salient fragments such as assessments, plans, deterioration, treatment response, and uncertainty statements. These fragments are then aligned with surrounding trajectories using rule-based windows informed by the earlier Condition Graphs and Physiology Templates. A subset of episodes can be manually reviewed to audit whether the chosen windows preserve the intended temporal and evidential interpretation.

This episode-based design keeps benchmark instances focused enough for transparent error analysis while avoiding collapse into coarse admission-level labels, which often hide whether a model relied on the right evidence or merely exploited dataset shortcuts. In practice, the same patient may contribute multiple episodes corresponding to onset, escalation, response, or recovery, thereby exposing richer temporal reasoning patterns than a single downstream endpoint.

3.4 Task synthesis, adjudication, and failure analysis

The final stage turns benchmark episodes into reasoning tasks. The evaluation layer extends beyond a conventional static test set: task generation, answer keys, and evidence traces are all tied back to the episode definition. Rather than asking only whether a model predicts the right label, TIMELY-Agent asks whether the model can justify a clinically plausible conclusion using the right evidence at the right time. This emphasis is aligned with broader calls for clinically meaningful explanation and end-use evaluation in healthcare AI [18].

Initial tasks for **TIMELY-Bench** will probe temporal grounding, numeric and trend reasoning, threshold and transition detection, note-trajectory consistency, contrastive inference across neighbouring windows, and evidence attribution. These categories are intentionally diagnostic rather than leaderboard-oriented. A useful benchmark should not only rank models; it should reveal recurring failure modes such as temporal inversion,

over-reliance on one modality, unsupported causal language, or confident answers that cite irrelevant evidence.

Evaluation can also be layered. Some tasks can be derived semi-automatically from benchmark schemas and retrieval traces, while others can be clinician-checked on smaller audited subsets. This creates a practical compromise between scale and fidelity: fully manual curation is too expensive to sustain, but fully automatic task generation risks encoding errors or spurious assumptions. TIMELY-Agent therefore treats auditability as part of the benchmark contribution rather than an afterthought.

4 Pilot Scope, Work Packages, and Expected Outputs

To keep the project realistic for an early-stage workshop submission, the pilot will focus on a small set of exemplar acute care conditions with distinct temporal signatures, such as acute kidney injury, respiratory deterioration including ARDS-like trajectories, and fluctuating neurocognitive states. The aim is not to maximise disease coverage immediately, but to test whether the workflow can support acute, progressive, and fluctuating patterns within one unified benchmark design. These conditions are methodologically useful because they differ in anchoring logic: some rely on abrupt physiological change, some on evolving treatment response, and some on disagreement or complementarity between narrative and numeric evidence.

The pilot is organised into three work packages. **WP1** develops condition schemas, inclusion logic, and physiology templates through agent-assisted literature and guideline synthesis. **WP2** implements OMOP-native cohort retrieval, note chunking, and multimodal alignment to produce benchmark episodes with preserved provenance. **WP3** instantiates TIMELY-Bench tasks and runs baseline evaluations with representative clinical or general-purpose language models to characterise common reasoning failure modes. Across all work packages, the emphasis is on transparent methodology rather than end-to-end automation claims.

We expect four immediate outputs: a documented framework for agentic benchmark construction over standardised EHR data; a pilot benchmark of temporally aligned multimodal reasoning episodes; an initial evaluation suite targeting time-aware clinical reasoning rather than only static prediction; and a reusable methodological scaffold for future fellowship applications on trustworthy multimodal foundation models. Because the framework is standards-aware but not overly tied to one site-specific schema, it also creates a path for later extension, including FHIR-based upstream ingestion, broader OMOP deployment, and MEDS exports for downstream model training.

5 Relevance to the Field

TIMELY-Agent contributes to an increasingly important but underdeveloped area in health AI: the construction of evaluation resources that reflect how clinical reasoning actually unfolds across time and modalities. Its novelty lies less in proposing yet another model architecture and more in making the *benchmark construction process* itself explicit, modular, and standards-aware. This matters because it improves scientific transparency, practical portability across secure data environments, and evaluation quality: instead of asking only whether a model predicts the right label, TIMELY-Bench asks whether the model can justify conclusions using the right evidence at the right time.

As an ongoing PhD/fellowship-track research programme, this project is intentionally positioned between methodological pilot work and longer-term translational ambition. The workshop setting is therefore ideal: feedback can help refine benchmark scope, task granularity, module boundaries, adjudication strategy, and standards usage before expansion into a larger fellowship proposal. In that sense, TIMELY-Agent is both a concrete benchmark construction framework and a wider research agenda for evaluating multimodal clinical reasoning with greater rigour.

References

[1] Karan Singhal et al. “Large language models encode clinical knowledge”. In: *Nature* 620.7972 (2023), pp. 172–180. DOI: 10.1038/s41586-023-06291-2.

[2] Zeljko Kraljevic et al. “Foresight-a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study”. In: *The Lancet Digital Health* 6.4 (2024), e281–e290. DOI: 10.1016/S2589-7500(24)00018-8.

[3] Julián N. Acosta et al. “Multimodal biomedical AI”. In: *Nature Medicine* 28.9 (2022), pp. 1773–1784. DOI: 10.1038/s41591-022-01981-2.

[4] Shih-Cheng Huang et al. “Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines”. In: *npj Digital Medicine* 3.1 (2020), p. 136. DOI: 10.1038/s41746-020-00341-z.

[5] Alistair E. W. Johnson et al. “MIMIC-III, a freely accessible critical care database”. In: *Scientific Data* 3 (2016), p. 160035. DOI: 10.1038/sdata.2016.35.

[6] Alistair E. W. Johnson et al. “MIMIC-IV, a freely accessible electronic health record dataset”. In: *Scientific Data* 10.1 (2023), p. 1. DOI: 10.1038/s41597-022-01899-x.

[7] Swaraj Khadanga et al. “Using Clinical Notes with Time Series Data for ICU Management”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 6432–6437. DOI: 10.18653/v1/D19-1678.

[8] Ryan King, Tianbao Yang, and Bobak Mortazavi. “Multimodal Pretraining of Medical Time Series and Notes”. In: *Proceedings of Machine Learning for Health*. Vol. 225. Proceedings of Machine Learning Research. 2023, pp. 243–257.

[9] Marzyeh Ghassemi et al. “A review of challenges and opportunities in machine learning for health”. In: *AMIA Summits on Translational Science Proceedings* 2020 (2020), pp. 191–200.

[10] Alvin Rajkomar et al. “Scalable and accurate deep learning with electronic health records”. In: *npj Digital Medicine* 1.1 (2018), p. 18. DOI: 10.1038/s41746-018-0029-1.

[11] Hrayr Harutyunyan et al. “Multitask learning and benchmarking with clinical time series data”. In: *Scientific Data* 6 (2019), p. 96. DOI: 10.1038/s41597-019-0103-9.

[12] George Hripcsak et al. “Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers”. In: *Studies in Health Technology and Informatics* 216 (2015), pp. 574–578. DOI: 10.3233/978-1-61499-564-7-574.

[13] OHDSI. *OMOP Common Data Model*. URL: https://ohdsi.github.io/CommonDataModel/ (visited on 03/13/2026).

[14] Joshua C. Mandel et al. “SMART on FHIR: a standards-based, interoperable apps platform for electronic health records”. In: *Journal of the American Medical Informatics Association* 23.5 (2016), pp. 899–908. DOI: 10.1093/jamia/ocv189.

[15] HL7 International. *FHIR Overview*. URL: https://www.hl7.org/fhir/overview.html (visited on 03/13/2026).

[16] Medical Event Data Standard. *What is MEDS?* URL: https://medical-event-data-standard.github.io/docs/intro_pages/what_is_MEDS/ (visited on 03/13/2026).

[17] fastOMOP. *OMCP: Model Context Protocol Server for the OMOP Common Data Model*. URL: https://github.com/fastomop/omcp (visited on 03/13/2026).

[18] Sana Tonekaboni et al. “What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use”. In: *Proceedings of Machine Learning Research*. Vol. 106. 2019, pp. 359–380.

Adverse Events and Geriatric Syndromes in MIMIC-IV: A Multilabel Document Classification Study

Fahrurrozi Rahman[1], Aryo Pradipta Gema[2], Arlene Casey[3], Honghan Wu[4], Bruce Guthrie[1], and Beatrice Alex[1]

[1]Advanced Care Research Centre, University of Edinburgh, United Kingdom
[2]Institute for Language, Cognition and Communication, University of Edinburgh, United Kingdom
[3]Usher Institute, School of Population Health Sciences, University of Edinburgh, United Kingdom
[4]School of Health and Wellbeing, University of Glasgow, United Kingdom

1 Introduction

Adverse Events (AE) and Geriatric Syndromes (GS) are important clinical phenomena associated with increased morbidity, functional decline, and mortality in older adults [1, 2, 3]. Identifying these conditions in electronic health records (EHRs) is valuable for clinical research and patient safety monitoring, yet much of the relevant information is embedded in unstructured clinical narratives.

Recent advances in large language models (LLMs) have introduced new strategies for clinical NLP, including in-context learning (ICL) [4] and parameter-efficient fine-tuning (PEFT) such as low-rank adaptation (LoRA) [5]. This study investigates these approaches for identifying AE and GS in an annotated subset of the MIMIC-IV dataset.

2 Methods and Data

Dataset. We use a subset of discharge summaries from the MIMIC-IV database of patients aged 65 years or older. The dataset contains 2,200 manually double annotated documents and is partitioned at the document level into training/development (1,801 documents) and independent test sets (399 documents)[1].

The AE task comprises 28 labels (14 events and their negated counterparts) [6], while the GS task contains 24 labels (12 syndromes and their negated labels) [7]. Explicit negation is modelled as a separate label to distinguish explicit denial from simple non-mention.

Models and Experimental Setup. We conducted three modelling approaches for multilabel document classification: (1) ICL, (2) PEFT with LoRA, and (3) fine-tuning of a biomedical encoder-based transformer.

For ICL, we evaluated five instruction-tuned LLMs (1–8B parameters) from Llama (Llama 3.1-8B, Llama 3.2-1B, Llama 3.2-4B) [8], Phi-3-mini-128k [9], and Qwen3-4B [10]. Prompts specified the document-level multilabel task, the allowed labels and the output format. Three prompt variants were tested, differing in whether label definitions and additional rules were included. The number of in-context examples, k, ranged from 0 to 32 ($k \in \{0, 2, 4, 8, 16, 32\}$). Examples were selected either by using the same samples for

[1]The dataset will be released accompanied by a data paper.

all prediction texts or by performing similarity-based retrieval between the predicted text embeddings and the document embeddings from the training set using bioclinical-modernbert-base-embeddings[2].

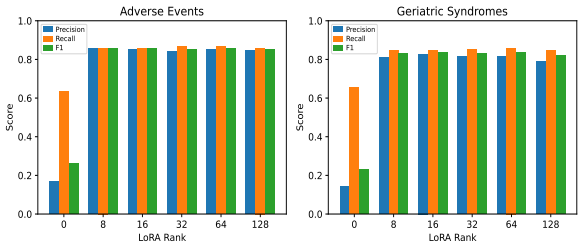
We first evaluated candidate models using ICL to estimate their task performance without training. Based on the resulting F1-score, Qwen3-4B was selected for PEFT. Multiple LoRA ($r \in \{8, 16, 32, 64, 128\}$) were evaluated to examine the trade-off between adaptation capacity and performance.

As a non-LLM baseline, we fine-tuned a biomedical encoder-based transformer (i.e., the bioclinical-modernbert-base-embeddings) using three strategies: full fine-tuning (Full), partial layer fine-tuning (Partial), and classification head only fine-tuning (Classification).

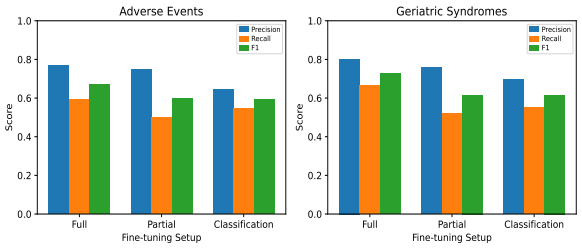
3 Results

Preliminary ICL experiments showed moderate performance and sensitivity to prompt design and the number of in-context examples. F1-score for prompts including label definitions produced more stable predictions (AE: 0.50, GS: 0.47), while similarity-based example selection slightly improved results (AE: 0.56, GS: 0.59).

For fine-tuning Qwen3-4B-Instruct, Figure 1a shows that performance improves with increasing LoRA rank up to a moderate value, after which gains plateau (F1-score AE: 0.86, GS: 0.83). Fine-tuning also substantially outperforms the baseline (F1-score: AE: 0.26, GS: 0.23). Figure 1b shows moderate performance for bioclinical-modernbert-base-embeddings, with full fine-tuning achieving the best results (F1-score AE: 0.67, GS: 0.73).



(a) Qwen3-4B-Instruct fine-tuning performance. Rank 0 is the baseline without fine-tuning.



(b) bioclinical-modernbert-base-embeddings fine-tuning performance.

Figure 1: Fine-tuning performance across models and configurations.

4 Conclusion

This study investigated approaches for multilabel classification of AE and GS from clinical discharge summaries. PEFT of Qwen3-4B-Instruct using LoRA showed that moderately sized LLMs can be effectively adapted for clinical multilabel classification tasks. Experiments with bioclinical-modernbert-base-embeddings showed that full fine-tuning performed best among the evaluated strategies. Future work will investigate factors that may explain remaining performance differences, including label imbalance, potential dependencies between labels, and error patterns accros event and syndrome categories.

[2] https://huggingface.co/NeuML/bioclinical-modernbert-base-embeddings

5 Study context

We conducted this study using a subset of the MIMIC-IV-Note dataset (MIMIC-IV) database [11].[3] Data access was approved by PhysioNet after completion of the CITI 'Data or Specimens Only Research' training. Prior to this study, A.C. conducted patient and public involvement and engagement (PPIE) activities related to the clinical NLP work within Advance Care Research Centre (ACRC), consulting patients on the use of free-text clinical data for AI research in healthcare and on the task of detecting AE and GS.

Institutional ethical approval for project this research was conducted in was granted by the Research Ethics Committee of B.A.'s host institution on 22/02/2022.

The work was conducted within ACRC, funded by Legal & General, and as part of the AIM-CISC project funded by NIHR (NIHR202639). F.R., H.W., B.G., and B.A. have been supported by Legal & General as part of the Advanced Care Research Centre (ACRC). A.P.G. is supported by the United Kingdom Research and Innovation (EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. A.C. is funded by the Vivensa Foundation (PF2302\2).

References

[1] Inouye SK, Studenski S, Tinetti ME, Kuchel GA. Geriatric Syndromes: Clinical, Research, and Policy Implications of a Core Geriatric Concept. Journal of the American Geriatrics Society. 2007;55(5):780-91. Available from: https://agsjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-5415.2007.01156.x.

[2] Long SJ, Brown KF, Ames D, Vincent C. What is known about adverse events in older medical hospital inpatients? A systematic review of the literature. International Journal for Quality in Health Care. 2013 10;25(5):542-54. Available from: https://doi.org/10.1093/intqhc/mzt056.

[3] Geyskens L, Jeuris A, Deschodt M, Van Grootven B, Gielen E, Flamaing J. Patient-related risk factors for in-hospital functional decline in older adults: A systematic review and meta-analysis. Age and Ageing. 2022 02;51(2):afac007. Available from: https://doi.org/10.1093/ageing/afac007.

[4] Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nature Medicine. 2024 Apr;30(4):1134-42. Available from: https://www.nature.com/articles/s41591-024-02855-5.

[5] Tran H, Yang Z, Yao Z, Yu H. BioInstruct: instruction tuning of large language models for biomedical natural language processing. Journal of the American Medical Informatics Association. 2024 06;31(9):1821-32. Available from: https://doi.org/10.1093/jamia/ocae122.

[6] Guellil I, Andres S, Anand A, Guthrie B, Zhang H, Hasan A, et al. Adverse Event Extraction from Discharge Summaries: A New Dataset, Annotation Scheme, and Initial Findings. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics; 2025. p. 28532-62. Available from: https://aclanthology.org/2025.acl-long.1386/.

[7] Guellil I, Andres S, Anand A, Guthrie B, Rahman F, Hasan AKMR, et al.. Geriatric Syndromes Extraction from Discharge Summaries: A New Dataset, Annotation Scheme and Initial Findings; 2026. Manuscript under review.

[3] https://physionet.org/content/mimic-iv-note/2.2/

[8] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al.. LLaMA: Open and Efficient Foundation Language Models; 2023. Available from: https://arxiv.org/abs/2302.13971.

[9] Abdin M, Aneja J, Awadalla H, Awadallah A, Awan AA, Bach N, et al.. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone; 2024. Available from: https://arxiv.org/abs/2404.14219.

[10] Yang A, Li A, Yang B, Zhang B, Hui B, Zheng B, et al.. Qwen3 Technical Report; 2025. Available from: https://arxiv.org/abs/2505.09388.

[11] Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. Mimic-iv. PhysioNet Available online at: https://physionet org/content/mimiciv/10/(accessed August 23, 2021). 2020.

TRExt: Demonstrating text analytics capabilities for Trusted Research Environments

Thomas Rowlands[1], Yamiko Msosa[2], Claire Newman[3], Philip Quinlan[1], Angus Roberts[2], Robert Stewart[2], Simon Thompson[3], Grazziela Figueredo[1], Tim Beck[1]
[1] University of Nottingham, Nottingham, UK
[2] King's College London, London, UK
[3] Swansea University, Swansea, UK

Introduction

Trusted Research Environments (TREs) are used to support secure access to sensitive health data for research while maintaining privacy protection and public trust. In parallel, federated analytics approaches are emerging as a powerful way to enable collaborative analysis across multiple TREs without requiring data to leave the secure environments in which they are held. These infrastructures have the potential to unlock large-scale population health insights while preserving institutional control over sensitive datasets. Current federated analytics capabilities are largely limited to structured datasets. A substantial proportion of clinically meaningful information remains embedded in unstructured text such as free-text clinical notes and correspondence. These textual records often contain rich descriptions of symptoms, outcomes, and social context that are not captured in structured fields. Despite their research value, such data are difficult to access at scale because they require specialised natural language processing (NLP) methods to extract the data, and anonymisation to make the data safe for analysis in TREs.

Unlocking the value of clinical text therefore requires approaches that can securely extract relevant clinical concepts and transform them into standardised, analytics-ready datasets suitable for large-scale research. While tools exist for clinical text mining and for transforming health data into standardised data models, these capabilities are not currently integrated within federated TRE infrastructures. This work introduces TRExt, a new capability to support Five Safes TES [1], a platform that enables federated analytics across TREs. TRExt integrates NLP and data mapping pipelines to convert unstructured clinical text into structured datasets aligned with widely used health data standards, the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) and the Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM). By enabling extraction, standardisation, and federated analysis of clinical text within TREs, this framework aims to expand the range of health data that can be reused for research within TREs while maintaining privacy, governance, and interoperability across institutions.

Methods and Data

Existing tools for clinical text processing, concept extraction, and data standardisation are integrated into a pipeline that generates safe data that can be analysed using federated analytics environments, such as Five Safes TES. Clinical information extraction is performed using two complementary NLP platforms. CogStack provides an extract–transform–load system integrating NLP pipelines capable of extracting medical concepts across multiple clinical domains. Additionally, the Mental Health Text Analytics Cloud (MH-TAC) platform is used to extract relevant concepts from psychiatric EHR notes and reports. Both systems utilise MedCAT for named entity recognition [2] and RelCAT for relation extraction [3], enabling identification of diseases, symptoms, medications, and contextual attributes within text.

Extracted clinical concepts are mapped to standard vocabularies to enable interoperability across datasets and institutions. Lettuce, an AI-assisted mapping tool that combines lexical search and large language model support, is used to suggest mappings between extracted terms and OMOP vocabulary concepts [4]. Carrot, a software tool for automating the conversion of healthcare datasets to the OMOP CDM through rule-based data transformation workflows [5], then executes the transformation of the extracted and mapped data into OMOP.

To support interoperability between research and clinical trial data standards, the Unison data virtualisation platform is used to implement mappings between OMOP and CDISC SDTM. Rather than physically duplicating datasets, Unison enables virtual mapping between data models, allowing federated analyses written for CDISC datasets to be executed against OMOP-aligned data.

Results

Initial results have come from processing free-text discharge summaries from the MIMIC IV dataset. A Postgresql database contains OMOP "Note" and "Note_NLP" tables that are used to describe the source text and extracted entities respectively. The Note table was populated with 331,732 note records from the MIMIC dataset and used as input for the CogStack NLP pipeline. The MedCAT v2 Snomed2025 model was used to extract 631,789 entities from the text and normalise these with SNOMED CT codes. The output from the pipeline was used to populate the Note_NLP table. The source SNOMED CT codes are mapped to OMOP concepts using Carrot and Lettuce, followed by transformation into the OMOP CDM 5.4 clinical data tables.

Conclusion

This ongoing work presents TRExt, a technical framework that supports the Five Safes TES federated analytics platform to enable the secure processing and standardisation of clinical text within TREs. By integrating established clinical NLP tools with AI-assisted concept mapping and data transformation pipelines, TRExt enables information extracted from unstructured health records to be represented within widely adopted health data standards such as OMOP and CDISC. Enabling these capabilities for TRE providers has the potential to substantially expand the types of health data that can be used in multi-institutional research, improving the reuse of routine clinical data and supporting new forms of population health, clinical, and translational research. Future work will include evaluating the quality of the generated OMOP data using the OHDSI Data Quality Dashboard, improving the performance of the pipeline, and validating it with real sensitive data. We will also continue our ongoing engagement activities with researchers and the public to ensure responsible and trustworthy use of AI-enabled health data technologies.

Study context

This study forms part of the DARE UK Next-Gen Catalysts Programme and incorporates public involvement and engagement (PIE) activities coordinated through the SAIL Databank. Ethical approval for the PIE activities has been granted from the University of Nottingham REC. The project team includes collaborators from academic institutions, national data infrastructures, and health data research initiatives. The tools developed in this work will be released in a form that can be deployed by TRE providers to support secure, federated analysis of free-text data.

References

1. HDR UK Federated Analytics programme. Five Safes TES. GitHub. 2026; https://github.com/SwanseaUniversityMedical/5s-Tes.
2. Kraljevic Z, et al. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. Artif Intell Med. 2021;117:102083
3. Agarwal S, et al. RelCAT: Advancing Extraction of Clinical Inter-Entity Relationships from Unstructured Electronic Health Records. arXiv preprint 2025;arXiv:2501.16077.
4. Mitchell-White J, et al. Llettuce: An Open Source Natural Language Processing Tool for the Translation of Medical Terms into Uniform Clinical Encoding. arXiv preprint 2024;arXiv:2410.09076.
5. Cox A, et al. Conversion of Sensitive Data to the Observational Medical Outcomes Partnership Common Data Model: Protocol for the Development and Use of Carrot. JMIR Res Protoc. 2025;14:e60917.

A Neuro-Symbolic Approach to Graph-Verified and Interpretable Chest X-Ray Report Generation

Faezeh Safari, Hang Dong, Zeyu Fu, and Aline Villavicencio

University of Exeter, Devon, England, UK

1 Introduction

Automated chest X-ray report generation holds significant promise for reducing the reporting burden on radiologists and improving turnaround times in clinical settings. However, current neural systems still produce factual hallucinations — clinically erroneous assertions that undermine diagnostic reliability and remain a critical barrier to safe deployment [1, 2]. Recent transformer-based and large language model (LLM) approaches have achieved impressive lexical fluency, yet they function primarily as statistical text completers, relying on pattern recognition without genuine understanding of medical structure or causal-temporal relationships [3, 4]. Knowledge-guided methods partially address this by improving interpretability [5]. Following research introduces **Nesy-Gen** (Neuro-Symbolic Graph-guided Generator), a dual-path architecture that unifies visual perception, temporal biomedical knowledge graph reasoning, and ante-hoc verification. The system is designed to produce reports that are both linguistically coherent and independently verifiable against structured biomedical knowledge, offering a transparent and auditable evidence trail for clinical review.

2 Methods and Data

Datasets. We use two standard radiology benchmarks: IU X-ray [6] and MIMIC-CXR [7]. The 7:1:2 train/validation/test split is used for IU X-ray, following prior work [8], and the official split is applied for MIMIC-CXR. To prevent data leakage, all Chest ImaGenome longitudinal pairs used in temporal graph construction are strictly excluded from the test evaluation [9]. As illustrated in Figure 1, Nesy-Gen comprises three tightly coupled modules:

1. *Vision-Entity Cross-Attention (VECA).* A Swin Transformer encodes the chest X-ray into spatial patch features. Clinical indication text is parsed using ScispaCy [10] to extract medical entities, which are then embedded via TransE [11] and aligned to PrimeKG nodes. Cross-attention between image regions and indication entities ensures that visual representations are semantically grounded — for example, suppressing peripheral image regions when the clinical indication is "shortness of breath" in favour of pulmonary and cardiac regions. TransE is selected for its efficiency on large-scale KGs; entity embeddings are

frozen to preserve pre-learned relational structure, while projection layers remain trainable.

2. *Temporal Subgraph Construction.* For each patient, a subgraph of PrimeKG is extracted by identifying relevant disease, phenotype, anatomy, drug, and biological process nodes from the report and linking them via a temporal Steiner tree algorithm. Edge weights balance temporal distance and clinical relevance, prioritising causal and localisation relations. Synthetic temporal progression edges are generated from domain templates (e.g., `cardiomegaly` → `pulmonary oedema` → `pleural effusion`) to densify sparse longitudinal supervision.

3. *Multimodal Constraint Verification (Consistency Gate).* Before any token is accepted for output, it must satisfy three conditions simultaneously: (a) *NLI entailment* — a BioBERT-based verifier confirms the candidate statement is entailed by the visual evidence; (b) *visual grounding* — cross-attention confirms the predicted finding is spatially supported in the image; and (c) *graph reachability and logic consistency* — Logic Tensor Networks (LTN) [12] evaluate candidate assertions against first-order clauses encoding biological plausibility, anatomical localization validity, and finding-to-diagnosis connectivity. Tokens failing any condition are flagged or rejected; uncertain assertions receive uncertainty markers rather than silently propagating.

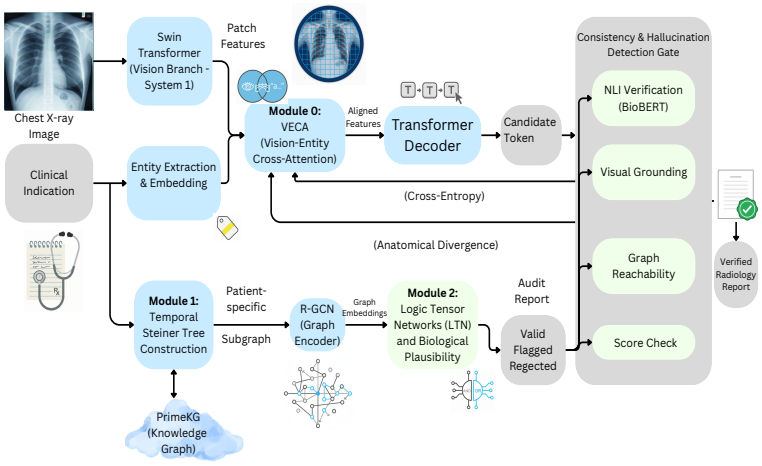


Figure 1: Nesy-Gen dual-path architecture. The Vision Branch (left) extracts spatially grounded image features aligned to clinical indications. The Neuro-Symbolic Branch (right) extracts a temporal patient subgraph from PrimeKG and evaluates logical consistency via LTN. The Consistency Gate (centre) accepts tokens only when entailment, visual grounding, and graph reachability are jointly satisfied.

Training and Hyperparameter Selection. Gate thresholds and loss weights are selected on validation data using a grid search that filters candidates by clinical safety constraints before optimising for a Pareto-optimal configuration balancing fluency and clinical faithfulness. Optimisation uses AdamW with a 5-epoch linear warmup followed by cosine decay. This staged approach ensures the final model prioritises biological consistency over raw fluency.

Evaluation Metrics. We report BLEU-1 through BLEU-4, ROUGE-L, and METEOR for text quality. Entity-level clinical correctness is assessed via Precision, Recall, and F1: for each predicted/reference report pair, medical entities are extracted and linked to PrimeKG nodes, and overlap is computed at the entity level. Per-sample scores are averaged across the test set to measure clinical concept recovery beyond surface fluency.

3 Results

Table 1 summarises performance on both datasets. On IU X-ray, Nesy-Gen achieves state-of-the-art results across most metrics including BLEU and entity-level F1. On MIMIC-CXR, the model records the best METEOR score (27.4 vs. 16.7 for TRRG, the strongest competing baseline) and the highest clinical F1 (43.1 vs. 39.3), demonstrating that the neuro-symbolic constraints yield the most substantial gains on the more clinically complex dataset. Adding VECA alone improves visual-indication alignment, reflected in BLEU gains. The LTN module provides the largest single-component gain in clinical precision and F1, confirming that logical consistency verification is the most impactful component for entity-level correctness.

Table 1: Performance comparison on IU X-ray and MIMIC-CXR. **Bold**: best. Underline: second best. BL = BLEU, MR = METEOR, RGL = ROUGE-L.

Model / Variant	BL1	BL2	BL3	BL4	MR	RGL	P	R	F1
IU X-ray Dataset									
KGAE [13]	41.7	26.3	1.81	12.6	14.9	31.8	–	–	–
TRRG [14]	48.2	30.2	21.7	15.1	**20.9**	37.7	–	–	–
R2Gen [8]	47.0	30.4	21.9	16.5	18.7	37.1	–	–	–
Baseline (Swin+Trans)	39.2	24.5	17.8	11.2	15.5	31.2	33.8	27.4	30.3
+ VECA	44.1	29.8	21.1	15.4	18.8	34.5	37.1	31.8	34.2
+ Temporal ST	46.5	31.2	22.4	16.8	19.9	36.8	39.4	34.2	36.6
+ LTN	47.8	32.4	23.5	17.9	20.1	37.2	40.6	35.6	37.9
Ours (Full)	**48.8**	**33.4**	**24.1**	**18.5**	20.8	**37.9**	**41.2**	**36.5**	**38.7**
MIMIC-CXR Dataset									
KGAE [13]	22.1	14.4	9.6	6.2	9.7	20.8	–	–	–
TRRG [14]	**43.6**	**29.8**	**21.3**	**15.7**	16.7	**33.6**	40.3	**39.9**	39.3
R2Gen [8]	35.3	21.8	14.5	10.3	14.2	27.7	33.3	27.3	27.6
Baseline (Swin+Trans)	28.5	15.2	8.8	5.4	11.8	22.4	30.1	24.0	26.7
+ VECA	33.2	19.4	12.6	8.2	14.5	27.2	36.2	29.1	32.3
+ Temporal ST	35.8	21.0	13.8	9.9	22.0	28.5	41.0	33.4	36.8
+ LTN	36.8	21.8	14.4	10.2	26.8	29.4	44.7	35.9	39.8
Ours (Full)	37.5	22.8	14.1	9.8	**27.4**	30.2	**48.5**	38.8	**43.1**

Qualitative Analysis. Token-level inspection (Figure 2) confirms that clinically unsupported phrases are suppressed when entailment, visual grounding, or graph reachability fails. Subgraph visualisations (Figure 3) provide an auditable evidence trail — from extracted entities to accepted report statements — that supports both error analysis and clinician review.

4 Conclusion

This work demonstrates that integrating neuro-symbolic constraints into multimodal radiology report generation can meaningfully improve clinical reliability beyond what fluency-focused neu-

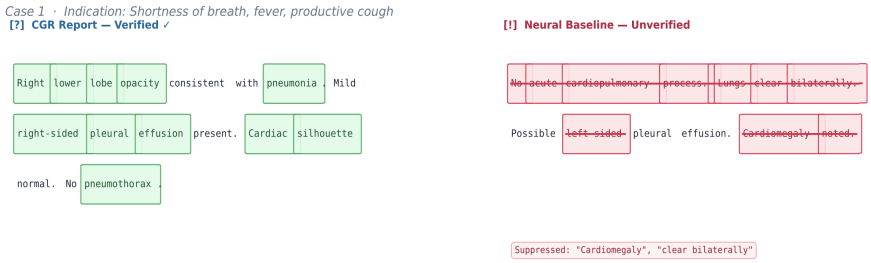


Figure 2: Token-level comparison of Nesy-Gen (left) and R2Gen [8] (right) on IU X-ray. Nesy-Gen suppresses baseline-only errors such as *cardiomegaly* ($s_{ground} = 0.18$) and *clear bilaterally* (no $\mathcal{T}^*$ path to `Normal Lung`).

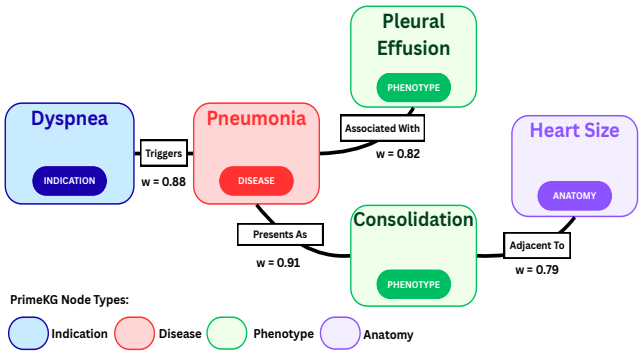


Figure 3: Temporal Steiner Subgraph extracted from PrimeKG for Dyspnea (IU-Xray).

ral baselines achieve. Nesy-Gen's ante-hoc verification approach — enforcing entailment, visual grounding, and biological plausibility jointly before token generation — strengthens entity-level correctness across two established benchmarks. Current results show a 64% relative METEOR improvement on MIMIC-CXR over the strongest prior baseline and the best clinical F1 score overall. Key limitations include dependence on KG coverage and the quality of manually curated temporal progression rules, which may not generalise to rare findings. Future research will focus on several directions to address these limitations. First, better uncertainty calibration is required. For example, using conformal prediction methods with LTN scores will help the Consistency Gate provide more detailed and accurate uncertainty markers. This is particularly useful for rare cases where PrimeKG data is limited. Second, we are interested to improve the temporal subgraph construction by including patient history from Chest ImaGenome. This will reduce the need for manual rules and help the model handle unusual disease changes. Third, the entity-level F1 metric is useful, but it does not show how serious a disease is. Therefore, future evaluations will combine RadGraph scoring and CheXpert label matching to better measure the clinical accuracy of the reports. Finally, the current model creates reports in one step. We could explore a step-by-step improvement method where the Consistency Gate asks the model to rewrite only the wrong parts to improve the report quality.

5 Study Context

This submission reports research in the Department of Computer Science, University of Exeter. The work is being developed in support of a planned studentship proposal. No patient contact or prospective clinical intervention is involved; all experiments use de-identified, publicly available datasets under their respective usage terms (IU X-ray [6], MIMIC-CXR [7], Chest ImaGenome [9]). Research focus is on transparent AI behaviour for clinical safety. No conflicts of interest are declared at this stage.

References

[1] Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers); 2018. p. 2577-86.

[2] Jing B, Wang Z, Xing E. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In: Proceedings of the 57th annual meeting of the association for computational linguistics; 2019. p. 6570-80.

[3] Wang Z, Liu L, Wang L, Zhou L. R2gengpt: Radiology report generation with frozen llms. Meta-Radiology. 2023;1(3):100033.

[4] Wang R, Xu Z, Wang X, Liu W, Lukasiewicz T. C2M-DoT: Cross-modal consistent multi-view medical report generation with domain transfer network. Information Fusion. 2026;125:103442.

[5] Safari F, Dong H, Fu Z, Villavicencio A. GraphRAG-Rad: Concept-Aware Radiology Report Generation via Latent Visual-Semantic Retrieval. In: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics; 2025. In press.

[6] Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, et al. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association. 2016;23(2):304-10.

[7] Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng Cy, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data. 2019;6(1):317.

[8] Chen Z, Song Y, Chang TH, Wan X. Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP); 2020. p. 1439-49.

[9] Wu JT, Agu NN, Lourentzou I, Sharma A, Paguio JA, Yao JS, et al. Chest imagenome dataset for clinical reasoning. arXiv preprint arXiv:210800316. 2021.

[10] Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP workshop and shared task; 2019. p. 319-27.

[11] Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. Advances in neural information processing systems. 2013;26.

[12] Badreddine S, Garcez Ad, Serafini L, Spranger M. Logic tensor networks. Artificial Intelligence. 2022;303:103649.

[13] Liu F, You C, Wu X, Ge S, Sun X, et al. Auto-encoding knowledge graph for unsupervised medical report generation. Advances in Neural Information Processing Systems. 2021;34:16266-79.

[14] Wang Y, Sun Y, Tan T, Hao C, Cui Y, Su X, et al. TRRG: Towards Truthful Radiology Report Generation With Cross-Modal Disease Clue Enhanced Large Language Models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2025. p. 647-57.

Tools for User-focused Mining of the Biomedical Literature

Neil R. Smalheiser[1]
[1] **University of Illinois-Chicago, Chicago, USA**

Introduction

PubMed is arguably the leading search engine for biomedical literature, encompassing well-curated articles indexed in MEDLINE, open access full-text articles indexed in PubMed Central, and other collections. A tremendous amount of informatics research has been devoted to improving public users' experience in retrieving articles relevant to a given query [e.g., 1-5]. Nevertheless, PubMed focuses on the broad needs of non-expert users, and offers limited options for detailed mining of retrieval results.

Our research team has created a suite of free, public web-based tools that piggyback on PubMed queries and allow users to conduct specialized searches and mining analyses. I will provide demonstrations of one or more of these tools in real time in response to queries suggested by conference attendees.

Methods and Data

All tools described here, together with the underlying models and data, have been fully described in previous publications and either derive from PubMed metadata or are freely available on our project website https://arrowsmith.psych.uic.edu.

Results

The following tools will be available for demo at the conference:

1. **ARROWSMITH** [6-8]. Identifies meaningful links between two sets of PubMed articles. This was the original tool devoted to Literature Based Discovery in 2001 and still maintained to date.

2. **Anne O'Tate** [9, 10]. Carries out a PubMed query and permits users to drill-down, expand, summarize and mine retrieval results in a variety of ways. Some of its functionalities are unique and not available anywhere else. For example, one can selectively retrieve articles that have a specific number of authors on the paper, or articles that have been cited a particular number of times. One can also view words and phrases that are particularly over-represented in the retrieval set, author names, affiliations, journals, publication dates, and various ways of analyzing topics discussed. Of note, the tool displays publication types and study designs of the retrieved articles, both those indexed by NLM and those predicted by our in-house tool, **Multi-Tagger** [11]. (Note that Multi-Tagger will soon be superceded by an improved transformer-based model not yet in production [12, 13]). For any article, the tool also displays the **Citation Cloud** that surrounds it, i.e., those articles that cite it, are cited by it, are co-cited with it, and that are bibliographically coupled to it [14].

3. **Trial-related tools**. Besides the tools that are displayed within Anne O'Tate, several standalone tools are devoted to retrieving and analyzing clinical trial articles. For example, **Trials to Publications** allows users to input one or more registered trials in ClinicalTrials.gov, and receive a list of articles (both explicitly linked and predicted) that are likely to arise from that trial [15-17]. **RCT Tagger** takes one or more articles as

input and estimates the probability that the article(s) describe randomized controlled trial outcomes [18]. **Aggregator** takes a set of trial articles as input and attempts to identify which articles derive from the same underlying trial [19, 20].

4. **Finding Case Report Nuggets**. This tool takes as input a topical PubMed query (e.g., bicycle injuries) and seeks to find “nuggets”, which are groups of five or more case reports that share the same or very similar main findings [21, 22]. This is intended to improve the signal-to-noise inherent in the case report literature.

Conclusion

Our suite of tools support a variety of specialized needs for individual users and research groups, ranging from identifying suitable reviewers for a given manuscript, to collecting trial and case report evidence for systematic reviews, to tracking how ideas flow through the literature, to assessing hypothesized links between findings made in disparate fields. A limitation is that our tools are based on PubMed and do not include all studies in all languages, nor in all ancillary fields such as pharmacy and physical therapy.

Study context

Ethics consideration and approvals – not applicable.

Funding -- all tools described here have been supported by NIH grants to N.R.S. Current support is R01 LM014292/LM/NLM NIH HHS/United States. Funder had no influence on the study, its design, or its publication.

Stakeholder involvement – several tools have been developed with strong input from user groups, particularly ARROWSMITH [23] and Multi-Tagger (submitted for publication).

Availability of data and methods -- https://arrowsmith.psych.uic.edu.

Conflicts of interest – none declared.

Collaborators – collaborators are co-authors on the references listed for each tool.

References

1. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70. doi: 10.1093/nar/gkh061.

2. McCray AT. The nature of lexical knowledge. Methods Inf Med. 1998 Nov;37(4-5):353-60.

3. Medical Subject Headings. https://www.nlm.nih.gov/mesh/meshhome.html

4. Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics. 2007 Oct 30;8:423. doi: 10.1186/1471-2105-8-423.

5. Krithara A, Mork JG, Nentidis A, Paliouras G. The road from manual to automatic semantic indexing of biomedical literature: a 10 years journey. Front Res Metr Anal. 2023 Sep 29;8:1250930. doi: 10.3389/frma.2023.1250930.

6. Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial intelligence. 1997 Apr 1;91(2):183-203.

7. Smalheiser NR, Swanson DR. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Comput Methods Programs Biomed. 1998 Nov;57(3):149-53. doi: 10.1016/s0169-2607(98)00033-9.

8. Smalheiser NR. Rediscovering Don Swanson: the past, present and future of literature-based discovery. Journal of Data and Information Science. 2017 Dec 29;2(4):43-64.

9. Smalheiser NR, Zhou W, Torvik VI. Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. J Biomed Discov Collab. 2008 Feb 15;3:2. doi: 10.1186/1747-5333-3-2.

10. Smalheiser NR, Fragnito DP, Tirk EE. Anne O'Tate: Value-added PubMed search engine for analysis and text mining. PLoS One. 2021 Mar 8;16(3):e0248335. doi: 10.1371/journal.pone.0248335.

11. Cohen AM, Schneider J, Fu Y, McDonagh MS, Das P, Holt AW, Smalheiser NR. Fifty ways to tag your pubtypes: Multi-tagger, a set of probabilistic publication type and study design taggers to support biomedical indexing and evidence-based medicine. medRxiv. 2021 Jul 16:2021-07.

12. Menke JD, Kilicoglu H, Smalheiser NR. Publication Type Tagging using Transformer Models and Multi-Label Classification. AMIA Annu Symp Proc. 2025 May 22;2024:818-827.

13. Menke JD, Ming S, Radhakrishna S, Kilicoglu H, Smalheiser NR. Enhancing automated indexing of publication types and study designs in biomedical literature using full-text features. medRxiv [Preprint]. 2025 Apr 28:2025.04.23.25326300. doi: 10.1101/2025.04.23.25326300.

14. Smalheiser NR, Schneider J, Torvik VI, Fragnito DP, Tirk EE. The Citation Cloud of a biomedical article: a free, public, web-based tool enabling citation analysis. Journal of the Medical Library Association: JMLA. 2022 Jan 1;110(1):103.

15. Smalheiser NR, Holt AW. A web-based tool for automatically linking clinical trials to their publications. Journal of the American Medical Informatics Association. 2022 May 1;29(5):822-30.

16. Holt AM, Troy AM, Smalheiser NR. Distribution of trial registry numbers within full-text of PubMed Central articles: implications for linking trials to publications and indexing trial publication types. Trials. 2025 Jan 31;26(1):34.

17. Holt AW, Smalheiser NR. Linking Trials to Publications: Enhancing Recall by Identifying Trial Registry Mentions in Full-Text. medRxiv [Preprint]. 2025 Jun 10:2025.06.09.25329285. doi: 10.1101/2025.06.09.25329285.

18. Cohen AM, Smalheiser NR, McDonagh MS, Yu C, Adams CE, Davis JM, Yu PS. Automated confidence ranked classification of randomized controlled trial articles: an aid

to evidence-based medicine. J Am Med Inform Assoc. 2015 May;22(3):707-17. doi: 10.1093/jamia/ocu025.

19. Shao W, Adams CE, Cohen AM, Davis JM, McDonagh MS, Thakurta S, Yu PS, Smalheiser NR. Aggregator: a machine learning approach to identifying MEDLINE articles that derive from the same underlying clinical trial. Methods. 2015 Mar;74:65-70. doi: 10.1016/j.ymeth.2014.11.006.

20. Smalheiser NR, Holt AW. New improved Aggregator: predicting which clinical trial articles derive from the same registered clinical trial. JAMIA Open. 2020 Oct 28;3(3):338-341. doi: 10.1093/jamiaopen/ooaa042.

21. Smalheiser NR, Shao W, Yu PS. Nuggets: findings shared in multiple clinical case reports. J Med Libr Assoc. 2015 Oct;103(4):171-6. doi: 10.3163/1536-5050.103.4.002.

22. Holt AW, Smalheiser NR. Finding Case Report Nuggets: A Web-Based Tool for Mining and Enhancing the Value of Clinical Case Reports. medRxiv [Preprint]. 2025 Nov 15:2025.11.13.25340162. doi: 10.1101/2025.11.13.25340162.

23. Smalheiser NR, Torvik VI, Bischoff-Grethe A, Burhans LB, Gabriel M, Homayouni R, Kashef A, Martone ME, Perkins GA, Price DL, Talk AC, West R. Collaborative development of the Arrowsmith two node search interface designed for laboratory investigators. J Biomed Discov Collab. 2006 Jul 3;1:8. doi: 10.1186/1747-5333-1-8.

Application of a BERT NLP model for recorded violence to investigate its associations with emergency department attendance and mental health service use in older adults.

Sharon Sondh[1], Christoph Mueller[1,2], Lifang Li[1,3], Angus Roberts[1], Harsharon Kaur Sondh[1], Robert Stewart[1,2]

[1]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[2]London and Maudsley NHS Foundation Trust, London, UK
[3] School of Journalism and Communication, Sun Yat-sen University, Guangzhou, Guangdong, PRC

Introduction: Abuse within older adults is a recognised global health concern (1) and can take many forms including physical, sexual, emotional or financial (2). Studies estimate that 10% of individuals over the age of 65% encounter some form of abuse (3). Abuse has been linked to a worsening of both psychological and physical health (1) and can result in co-occurring medical conditions and premature mortality (4). Specifically, abuse in older adults has been found to be associated with hospital use (5). In addition, abuse is often under recognised by emergency department professional resulting in a higher risk for further abuse(6).

Natural language processing has remained a valuable tool in processing and extracting valuable information from electronic health records (EHR). A previous NLP application was created which successfully extracted data on domestic, physical and sexual abuse (7). This study applies an updated version of this application which includes additional features such additional abuse types (psychological and financial) and the patient's role (8). Applying this NLP, this study aimed to investigate the associations between older adult abuse and use of mental health services and emergency department attendance.

Methods and Data: Data for this study was sourced using the Clinical Record Interactive Search (CRIS) to access deidentified EHR from the South London and Maudsley NHS Foundation Trust (SLaM). SLaM is one of the largest mental healthcare providers and covers four South East London boroughs. Data was sourced from both structured fields and free text using an NLP algorithm which builds on earlier work and used a fine-tune multi-label BERT model to train annotated data from 6,500 text instances. The model includes detection of physical, sexual, emotional and financial abuse. The current study deployed this application to extract a new dataset on abuse for older adults who had at least one face to face appointment in adult mental health services (at SLaM) between 1st January 2007 and 31st December 2022. The BERT model was trained further (using 95% of the annotated dataset) and the remaining used for blind testing. The performance varied according to the abuse type; capturing financial, emotional, sexual, physical abuse well (precision/recall estimations for 651 instances: 83%/95%,81%/95%, 78/91%, 89%/55%).

The outcomes assessed were emergency department attendance ascertained from linked Hospital Episode statistics and number of face-to-face contacts extracted via CRIS. Other characteristics extracted were age, gender, ethnicity, marital status and Index of Multiple Deprivation (IMD). IMD scores were divided into IMD tertiles. Ethnicity was dichotomised into White and non-White and marital status was dichotomised as married/cohabiting or not.

Analyses were carried out using STATA 18 software (9). Cox regression models were generated to examine the association between recorded abuse and emergency department attendance. Negative binomial regression models were run to assess the associations between recorded abuse and number of mental health service face-to-face contacts. Models were adjusted for age, gender, ethnicity, marital status and deprivation. The sample was then stratified by gender, ethnicity, marital status and deprivation to further analyse the differences in outcomes for recorded abuse types. Patients were followed up from index date to first emergency department attendance, date of death or censoring date (31st December 2022).

Results: Data from 14,591 patients seen by older adult mental health services was analysed. Patients had a mean age of 78.8 (7.5) years and 61.5% were female. 1,227 (8.4%) were experiencing abuse with physical abuse (64.6%, n=793) being the most common and 26.1% of these had records of more than one form of abuse. Those with recorded abuse were younger, more likely to be female, more likely to be married or cohabiting and from more deprived areas. The time until first emergency department attendance was found not to differ between those experiencing abuse (Survival time: 19.7, interquartile range: 8.6-47.6 months) and those not experiencing abuse (19.4, 8.7-43.7 months; p=0.284). In comparison to those without recorded abuse, the mean (SD) number of mental health contacts in the year was higher in those with recorded abuse (11.1 (11.5), p<0.001).

The regression model found no significant association between recorded abuse and emergency department attendance. A significant association between abuse and higher mental health service contact was found for those in a non-white ethnic group, patients married/cohabiting and those from the least deprived

neighbourhoods. Considering the types of recorded abuse, emotional abuse was found to be associated with a 13% increased risk of A&E attendance. All types of abuse were found to be significantly associated with higher mental health contacts with sexual abuse having the strongest associations (HR:2.35, CI:1.93-2.86).

Conclusion: Abuse experience is commonly recorded in older adults and has been linked to worsening physical and mental health (1). The current study applied a BERT-derived NLP algorithm successfully to investigate outcomes of recorded abuse, which would hitherto have not been possible using healthcare data. When considering abuse types, only emotional abuse was found to be associated with increased risk of emergency department attendance, but all types of abuse were associated with higher mental healthcare contacts.

The main constraints of the study were the restriction of the cohort being used from a single south London catchment area and some variable reduction for ethnicity and marital status when conducting the analysis. Further investigation would benefit from a broader geographical scope encompassing data from a larger spectrum. It would also be beneficial for further studies to highlight and raise awareness so that more vulnerable groups are captured.

Study context: The data analysed in this study is subject to the following licenses/restrictions: All the relevant aggregate data are found within the article. The data used in this work have been obtained from CRIS. It provides authorised researchers regulated access to anonymised information extracted from SLaM's electronic clinical records system. Individual-level data are restricted in accordance with the strict patient-led governance. Research use of the source data, including all work described here, is covered by approval from Oxford Research Ethics Committee C, reference 23/SC/0257. Data are available for researchers who meet the criteria for access to this restricted data: (i) SLaM employees or (ii) those having an honorary contract or letter of access from the trust.. Requests to access these datasets should be directed to CRIS administrator: cris.administrator@kcl.ac.uk.

Funding:
SS is a fully funded PhD student at the NIHR HealthTech Research Centre in Brain Health. AR is funded by the NIHR Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust. CM and RS are part-funded by the NIHR Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London and the NIHR HealthTech Research Centre in Brain Health. RS is additionally part-funded by (i) the NIHR Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust; (ii) the UK Research and Innovation (UKRI) – Medical Research Council through the DATAMIND HDR UK Mental Health Data Hub (MRC references: MR/W014386/1 and MR/Z504816/1); (iii) the UK Prevention Research Partnership (Violence, Health and Society; MR-VO49879/1), an initiative funded by the UK Research and Innovation Councils, the Department of Health and Social Care (England) and the UK devolved administrations and leading health research charities. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. HKS is a fully funded PhD student at the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre.

References:

1. Gagnon S, Nadeau A, Tanguay K, Archambault PM, Brousseau AA, Carmichael PH, et al. Prevalence and predictors of elder abuse among older adults attending emergency departments: a prospective cohort study. CJEM. 2023;25(12):953-8.
2. World Health Organisation. Abuse of older people 2024 [Available from: https://www.who.int/news-room/fact-sheets/detail/abuse-of-older-people.
3. Patel K, Bunachita S, Chiu H, Suresh P, Patel UK. Elder Abuse: A Comprehensive Overview and Physician-Associated Challenges. Cureus. 2021;13(4):e14375.
4. Alias AN, Mokti K, Ibrahim MY, Saupin S, Madrim MF. Elderly Abuse and Neglect on Population Health: Literature Review and Interventions from Selected Countries. Korean J Fam Med. 2023;44(6):311-8.
5. Rosen T, Zhang H, Wen K, Clark S, Elman A, Jeng P, et al. Emergency Department and Hospital Utilization Among Older Adults Before and After Identification of Elder Mistreatment. JAMA Netw Open. 2023;6(2):e2255853.
6. Mercier É, Nadeau A, Brousseau A-A, Émond M, Lowthian J, Berthelot S, et al. Elder abuse in the out-of-hospital and emergency department settings: a scoping review. Annals of emergency medicine. 2020;75(2):181-91.

7. Botelle R, Bhavsar V, Kadra-Scalzo G, Mascio A, Williams MV, Roberts A, et al. Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: an applied evaluative study. BMJ Open. 2022;12(2):e052911.
8. Li L, Sondh S, Sondh HK, Stewart R, Roberts A. Development of a natural language processing application to extract and categorize mentions of violence from mental healthcare records text. medRxiv. 2026:2026.03.22.26348435.
9. Noble M, McLennan D, Wilkinson K, Whitworth A, Exley S, Barnes H, et al. The English indices of deprivation 2007. 2007.

Clinical data enrichment using LLM ensemble approaches

Keiran Tait[1], Joseph Cronin[1], Robert Dürichen[1]

[1] Arcturis Data Ltd, Kidlington, Oxfordshire, UK

Introduction

Many clinically important details needed for high-quality real-world evidence (RWE) studies are captured only in unstructured clinical notes, including pathology and radiology reports. In this context, *clinical data enrichment* refers to the systematic transformation of such unstructured information into structured, analysis-ready datasets—improving completeness, consistency, and usability for downstream research rather than introducing new clinical knowledge. Large language models (LLMs) are increasingly used for this task, offering strong performance in interpreting complex clinical narratives and extracting structured variables [1], [2].

Despite these advantages, LLM outputs remain vulnerable to hallucinations and missing or defaulted values. Studies have shown that even guard-railed clinical pipelines require explicit mitigation strategies to minimise hallucinated or incomplete extractions [3], [4], [5]. Additionally, unlike smaller language models such as BERT-based systems [6], LLMs do not natively provide calibrated confidence metrics to indicate uncertainty in their outputs. A range of mitigation strategies has therefore been proposed, including layered prompting architectures, retrieval augmentation, post-generation verification, and ensemble approaches [7], [8].

These challenges are amplified in real clinical settings, where analysts often cannot access original reports owing to governance and confidentiality constraints. This limits direct verification of extracted values and makes it harder to assess transparency and output confidence in routine workflows.

In response, ensemble approaches have emerged as a practical strategy for improving reliability and making uncertainty more transparent. Surveys highlight ensemble LLMs as a growing research direction capable of improving robustness by reducing variance and correlated errors across models [8]. In operational pipelines, disagreement between ensemble members can be interpreted as a pragmatic proxy for extraction uncertainty, providing an auditable signal for prioritising human review rather than treating all LLM outputs as uniformly reliable. This paper evaluates three ensemble methods for extracting clinical features from pathology reports and quantifies their effect on both accuracy (benchmarking outputs against ground-truth annotations) and disagreement rate (the proportion of reports where ensemble members conflict). The aim is to identify practically deployable configurations that enrich structured datasets with reliable values while making uncertainty explicit, including configurations that achieve strong performance without always relying on the most computationally demanding models.

Methods and Data

Data: This study used 491 anonymised pathology reports from a single NHS trust. Five clinical features were analysed: oestrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), P53, and mismatch repair proteins (MMR). For each feature, 100 reports were manually annotated: 50 reports contained no mention of the feature or did not provide a specific value, while the remaining 50 included a clearly identifiable value (e.g., positive, negative, or other stated results).

Methodology: We evaluated three ensemble approaches for extracting clinical features from pathology reports:

- *Multi-prompt approach:* A single LLM is queried using two complementary prompts: (i) a feature-specific prompt using clinical context and few-shot examples, and (ii) a general clinical prompt that returns a structured JSON summary of all detected features. Outputs are compared to identify consistent values.
- *Multi-LLM approach:* Two different LLMs receive the same feature-specific prompt, allowing cross-model comparison. Disagreement highlights uncertain cases and reduces the risk of single-model hallucinations. This study focuses on two-model pairings.

- *LLM-as-a-judge approach:* An LLM first extracts a proposed value using a feature-specific prompt, then evaluates the plausibility of that extraction in relation to the full report via a second prompt using the same model.

Models: To assess the effect of model size and model families on ensemble behaviour, we evaluated a range of open-source instruction-tuned LLMs grouped into three parameter buckets: small (LLaMA-3.2-3B [9]), medium (Mistral-7B [10], LLaMA-3-8B [9], Qwen2-7B [11], Falcon-3-10B [12]), and large (Gemma-3-27B [13], LLaMA-3.1-70B [9], LLaMA-3.3-70B [9]).

Evaluation: We report extraction accuracy (ACC) for each individual model, and for ensemble approaches, the accuracy restricted to the subset of reports where ensemble members agreed (reported as agreement-filtered accuracy). We also report disagreement rate (DR), defined as the proportion of reports yielding conflicting outputs, representing cases that would require secondary validation. While DR is not a calibrated confidence estimate, it provides an interpretable proxy for extraction uncertainty in operational data pipelines.

Results

Our evaluation confirms the general trend that larger LLMs perform best: both LLaMA-3.1-70B and LLaMA-3.3-70B reached ~98% average accuracy, mid-sized models such as Falcon-3-10B (96.4%) and LLaMA-3-8B (96.2%) were close behind, while Qwen2-7B (92.0%) and Mistral-7B (87.2%) lagged, and LLaMA-3.2-3B performed weakest (73%).

Across ensemble strategies, the multi-prompt ensemble delivered the largest mean accuracy gain (+3.96%), but at the cost of a higher disagreement rate (18.78%)—with substantial variation by model (+0.9% accuracy for LLaMA-3.3-70B up to +13.72% for LLaMA-3.2-3B; disagreement rate from 8.6% for LLaMA-3.3-70B to 30.8% for Mistral-7B). By contrast, the LLM-as-a-judge approach yielded a smaller mean accuracy gain (+0.71%) but also a much lower disagreement rate (3.43%), though we note occasional accuracy decreases (e.g., −2.79% for LLaMA-3.2-3B; best gain +1.6% for Mistral-7B; disagreement spanning 1.2% for LLaMA-3-8B to 7.8% for Mistral-7B). For the two-LLM ensembles, efficacy depends strongly on the pairing. Using LLaMA-3-8B as an anchor, pairing with a 70B LLaMA raised accuracy by up to +2.35% with a moderate disagreement rate (~3%). Pairing models of the same size but different families was particularly promising with LLaMA-3-8B + Falcon-3-10B (Acc = 98.13%, with an accuracy gain for Llama-3-8B of +1.93%, 3.6% disagreement rate), whereas combinations with Qwen2-7B (+1.54%, 7.6%) and Mistral-7B (+0.71%, 18.6%) were less favourable.

Qualitative inspection of disagreement cases suggested common failure modes among weaker models, including confusion between equivocal and negative results, incorrect assumptions of absence when features were referenced indirectly, and hallucinated values when reports described pending or recommended tests. Larger models were more robust to fragmented reporting styles and long-range dependencies, and cross-family ensembles likely improved performance because their errors are less correlated than those of closely related models.

Conclusion

Our evaluation demonstrates substantial variation across both the tested LLMs and the ensemble techniques, with model size and architecture influencing extraction accuracy and disagreement rates. While larger models achieved the strongest standalone performance, carefully selected ensemble configurations—particularly cross-family medium-sized model pairs—can provide comparable accuracy while making extraction uncertainty explicit via model disagreement. Notably, combining LLaMA-3-8B with Falcon-3-10B produced high accuracy for a medium-sized model pair, with only 3.6% of reports flagged for further review, indicating a strong balance between reliability and operational efficiency. More broadly, this pattern is consistent with the general ensemble-learning principle that ensembles are most beneficial when member models do not fail in the same way. Recent evidence in LLM settings shows that error correlations can be substantial and are influenced by shared architectures and providers, motivating the use of heterogeneous model families when feasible [14]. These findings suggest that ensemble disagreement can serve as a transparent and operationally useful signal for prioritising human review, enabling more efficient and auditable clinical data enrichment workflows without reliance on the most computationally demanding models.

Study context

An anonymised dataset for this study was provided by the Arcturis Real-World Data Network research database under REC approval 24/YH/0164. Our thanks to all participating NHS trusts who provide data to the database, and to the patients and members of the public who advise and support Arcturis. Funding: This work was supported by internal funding from Arcturis Data Ltd (no external funding was received).

References

[1] L. Huang *et al.*, 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions', *ACM Trans. Inf. Syst.*, p. 3703155, Nov. 2024, doi: 10.1145/3703155.

[2] H. Ying *et al.*, 'GENIE: Generative Note Information Extraction model for structuring EHR data', Jan. 30, 2025, *arXiv*: arXiv:2501.18435. doi: 10.48550/arXiv.2501.18435.

[3] N. Dao *et al.*, 'Generative artificial intelligence for automated data extraction from unstructured medical text', *JAMIA Open*, vol. 8, no. 5, p. ooaf097, Sep. 2025, doi: 10.1093/jamiaopen/ooaf097.

[4] Y. Hu *et al.*, 'Improving large language models for clinical named entity recognition via prompt engineering', *J Am Med Inform Assoc*, vol. 31, no. 9, pp. 1812–1820, Sep. 2024, doi: 10.1093/jamia/ocad259.

[5] Y. Hu *et al.*, 'Information Extraction from Clinical Notes: Are We Ready to Switch to Large Language Models?', Jan. 07, 2025, *arXiv*: arXiv:2411.10020. doi: 10.48550/arXiv.2411.10020.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', May 24, 2019, *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.

[7] S. Hiriyanna and W. Zhao, 'Multi-Layered Framework for LLM Hallucination Mitigation in High-Stakes Applications: A Tutorial', *Computers*, vol. 14, no. 8, p. 332, Aug. 2025, doi: 10.3390/computers14080332.

[8] I. D. Mienye and T. G. Swart, 'Ensemble Large Language Models: A Survey', *Information*, vol. 16, no. 8, p. 688, Aug. 2025, doi: 10.3390/info16080688.

[9] Meta AI, 'LLaMA 3: Open Foundation and Instruction-Tuned Language Models', 2024.

[10] A. Jiang *et al.*, 'Mistral 7B', *arXiv*:2310.06825, 2023.

[11] J. Bai *et al.*, 'Qwen Technical Report', *arXiv*:2309.16609, 2023.

[12] N. Almazrouei *et al.*, 'The Falcon Series of Open Language Models', *arXiv*:2306.01116, 2023.

[13] Google DeepMind, 'Gemma: Open Models Based on Gemini Research and Technology', 2024.

[14] E. M. Kim, A. Garg, K. Peng, and N. Garg, 'Correlated Errors in Large Language Models', in *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. Available: https://proceedings.mlr.press/v267/kim25e.html

Cogstack Coder: Agentic Medical Coding Assistant for EHR Systems

Samuel Thio[1,2,3], David Tang[1,7], James Teo[4,5,7], Thomas Searle[1,7], Richard Dobson[1,6,7]
[1] Department of Biostatistics & Health Informatics, King's College London, London, UK.
[2] UKRI Engineering and Physical Sciences Research Council DRIVE-Health CDT, London, UK
[3] NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, London, UK
[4] Guy's and St Thomas' NHS Foundation Trust, London, UK
[5] Department of Neurology, King's College Hospital NHS Foundation Trust, London, UK
[6] Health Data Research UK London and Institute of Health Informatics, University College London, London, UK
[7] CogStack Limited, London, U.K

Introduction

Medical coding (translating clinical narratives into standardized ICD-10 and OPCS-4 codes) is the financial backbone of the NHS, directly determining how billions in funding are allocated to hospitals [1,2]. Every discharge summary must be reviewed by a trained clinical coder who reads dense free text, identifies relevant conditions and procedures, and assigns the correct codes. Yet the profession faces a convergence of crises: 1) The NHS vacancy rate for coders sits at 6.7% with over 100,000 unfilled posts as of late 2025; 2) Coding errors cost the NHS up to £1 billion annually in inaccurate reimbursement; 3) Coders often struggle with poor source documentation from clinicians [3].

Cogstack currently has deployed natural language processing models to perform medical coding and this has shown an uplift of £2 million per year annually per NHS Trust; however, long-stay complicated and impatient cases are not able to be solved with this current solution. There is a need for more robust handling of lengthy clinical records [4]. This project thus aims to explore the use of large language models and semi-autonomous agents in agentic architecture to be able alleviate some of the workflow pressures that medical coders face.

Methods and Data

The system is a multi-container application consisting of a Google ADK agentic backend (Python), a Clinical API service (FastAPI), and a frontend (TypeScript/React). Patient data is stored and searched via OpenSearch (MIMIC-IV dataset), code lookups use ChromaDB, and agent sessions persisted in SQLite with multi-tenant isolation. MedGemma 1.5 integration: MedGemma 1.5 is integrated at two levels (1) as the LangExtract extraction model for structured entity extraction from clinical text, and (2) as the Clinical API's LLM for code refinement and mapping [5]. The system supports multiple deployment modes for MedGemma 1.5: cloud via Gemini API, local via vLLM (quantized 4-bit), or via Ollama all configurable through environment variables without code changes.

Agentic workflow (via Google ADK): The ReACT-style agent [6] recursively searches EHR documents, extracts entities, maps to ICD-10 and OPCS-4, and presents results after completing a 5-step pipeline autonomously per discharge summary (see Figure 1).

The large language model entity extraction step uses Google's LangExtract package for smart chunking of documents with parallel parsing to assign various clinical attributes towards each entity extracted. Each subagent has session context injection via callbacks; and for robust agentic execution in production systems, our system has a system for error handling. There is a custom callback for tool calls to prevent hallucination of patients' and document ID by the agent and additionally uses an additional 'Reflect-and-Retry' tool to allow the agent to retrace their approach in case of errors.

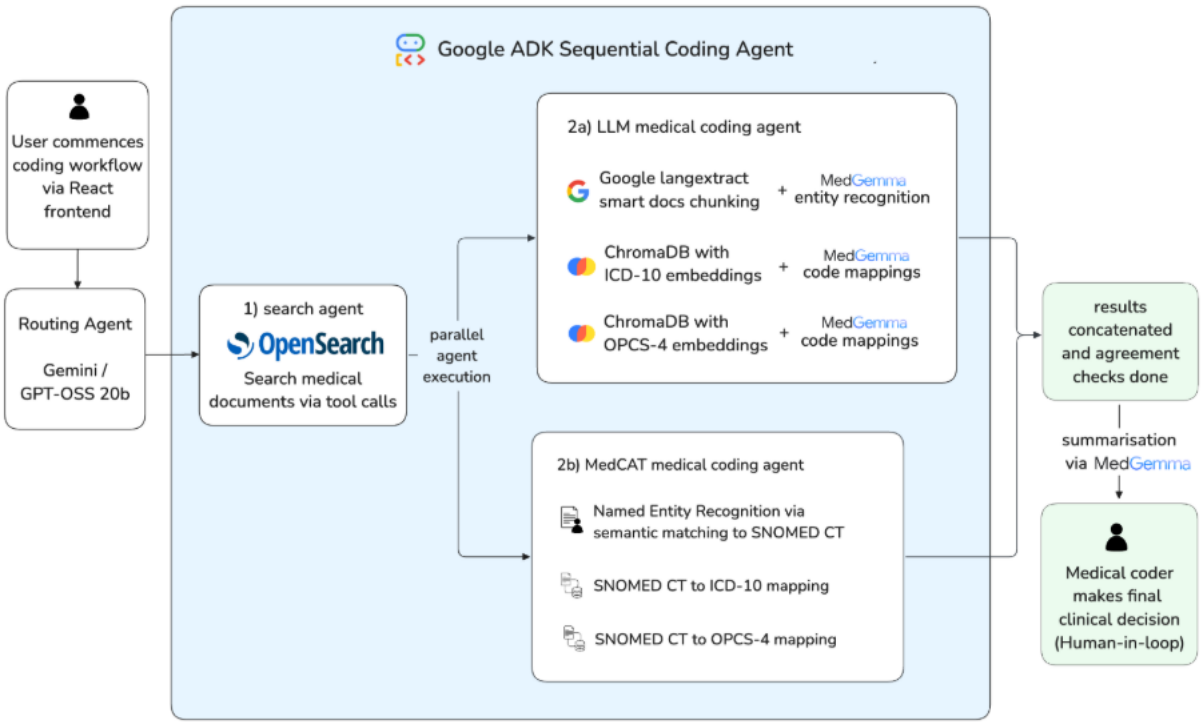


Figure 1. Agentic architecture displaying how ReACT subagents and agents are laid out in sequential order to accomplish the coding workflow. The icons denote the tools that are exposed to each subagent.

Data for Agent Grounding

A vector database of ICD-10 NHS (5th edition) and OPCS-4.11 codes and the app uses retrieval augmented generation (RAG) to ensure outputs are grounded to the NHS' required outputs. [1,2]

Deployment & feasibility

The application is deployed through two dockerized containers on the King's College London Computational Research, Engineering and Technology Environment and uses Langfuse observability tracing and PostHog analytics [7]. Open weight models are served via vLLM with document caching on an Nvidia A100 Tensor Core GPU. The privacy-first option allows the entire stack including MedGemma 1.5 to run locally with no data leaving the institution's network [7]. Key challenges and mitigations: LLM hallucination in medical coding is addressed through the dual-pipeline validation approach (compare MedGemma 1.5 output against deterministic MedCAT mappings) [8]. Clinical adoption requires trust addressed through traceable provenance from extracted text span to SNOMED CUI to final code, and a human-in-the-loop review interface where coders accept/reject each suggestion.

Conclusion

Cogstack Coder is a modern, on-prem agentic workflow system that uses a dual AI-system validation pathway for medical coding with a consideration for the human-in-loop workflow. Currently it is deployed on CREATE Cloud and being co-developed with the medical coding team at Guys and St Thomas' NHS Trust. Future evaluation and results work will be carried out across both the MIMIC-IV dataset and local NHS data.

Study context

There are no conflicts of interest to declare.

This work does not require ethics consideration as it is part of service improvement. Current iteration does not involve the use of any live patient data from NHS Trusts.

Collaborators: Guys and St Thomas NHS Trust medical coding team

Data availability: MIMIC-IV dataset is available to researchers via PhysioNet

References

1. NHS TRUD. NHS ICD-10 5th Edition data files. Accessed February 20, 2026. https://isd.digital.nhs.uk/trud/users/authenticated/filters/0/categories/28/items/258/releases
2. NHS TRUD. OPCS-4 data files. Accessed February 20, 2026. https://isd.digital.nhs.uk/trud/users/authenticated/filters/0/categories/10/items/119/releases
3. NHS Providers. NHS Digital workforce statistics – November 2025. Accessed February 23, 2026. https://nhsproviders.org/resources/nhs-digital-workforce-statistics-november-2025/
4. Case study update - using AI to unlock health records. 2024. Accessed March 8, 2026. https://www.hfma.org.uk/system/files/2024-09/CogStack%20AI%20-%20case%20study%20update%20v4.pdf
5. Sellergren A, Kazemzadeh S, Jaroensri T, et al. MedGemma technical report. *arXiv*. Preprint posted online July 12, 2025. doi:10.48550/arXiv.2507.05201
6. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, et al. ReAct: synergizing reasoning and acting in language models. *arXiv*. Preprint posted online 2023. https://arxiv.org/abs/2210.03629
7. King's Computational Research, Engineering and Technology Environment (CREATE). Accessed March 2, 2022. https://doi.org/10.18742/rnvf-m076
8. Kraljevic Z, Searle T, Shek A, et al. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. *Artif Intell Med*. 2021;117:102083. doi:10.1016/j.artmed.2021.102083

TIMELY-Bench: Quantifying Temporal Leakage in Multimodal ICU Prediction

Haoyu Wang[1], Zitong Li[1], Linglong Qian[1], and Zina Ibrahim[1]

[1]Department of Biostatistics and Health Informatics, King's College London, London, UK

1 Introduction

Multimodal prediction in intensive care increasingly combines structured physiological trajectories with clinical notes [1, 2, 3]. However, fair comparison across studies remains difficult because the two modalities follow different temporal recording processes: vital signs and laboratory tests are charted close to observation time, whereas notes are written more sparsely and often summarise events over a broader period. As a result, seemingly minor alignment choices can introduce post-anchor information and inflate predictive performance [4, 5]. Despite growing interest in multimodal ICU modelling, there is still no standard benchmark for evaluating how time-alignment protocols affect performance or for distinguishing leakage arising from future structured values versus future-oriented note content. This is particularly important in note-centred pipelines, where a note timestamp is used as the reference point for constructing multimodal inputs, but the note itself may contain retrospective and prospective statements.

We present TIMELY-Bench, a reproducible note-anchored benchmark for multimodal ICU prediction in MIMIC-IV [6]. The benchmark defines leakage-controlled alignment settings and uses a simple 2×2 analysis to separate apparent performance gains due to future structured measurements from those due to note content. Across four clinical prediction tasks, we show that AUROC inflation under leaked conditions is driven almost entirely by structured data leakage, while text leakage is negligible under note-level pooled embeddings.

2 Methods

Cohort and tasks. We used MIMIC-IV [6] to construct a multimodal ICU cohort with 74,829 stays. Each stay was represented by 42 structured variables extracted over the first 72 hours, together with clinical notes recorded within the first 48 hours. We evaluated two patient-level outcomes: in-hospital mortality and prolonged ICU length of stay. To test whether the same temporal effects generalise beyond standard benchmark tasks, we additionally examined progression from Acute Kidney Injury (AKI) Stage 1 to Stage 2 or above [7], and progression from sepsis to septic shock [8].

Note-anchored alignment. Each note timestamp T was treated as an anchor time for constructing multimodal inputs. For structured data, we defined several pre-anchor lookback windows, including same-day aggregation up to T (D0) and rolling 6-, 12-, and 24-hour windows (W6/W12/W24). To create an intentionally leaked setting, we also defined a symmetric 48-hour window spanning $[T-24, T+24]$, which includes post-anchor measurements. For the patient-level benchmark release, we retained the final note occurring within the first 48 hours of each ICU stay as the representative anchor instance.

Text representations and leakage control. Clinical notes within the selected text window were embedded using ClinicalBERT [3]. We considered two text conditions. In the *original text* condition, note embeddings were used without temporal filtering. In the *clean text* condition, we reduced the influence of future-oriented content using DocTimeRel labels, down-weighting notes with a higher proportion of sentences marked AFTER relative to document time [9].

Baselines and evaluation. We compared structured-only, text-only, early-fusion, and late-stacking baselines using logistic regression and XGBoost. Evaluation used predefined subject-level splits with a held-out test set and 5-fold cross-validation on the training portion.

2×2 leakage analysis. To quantify where performance inflation comes from, we focused on early-fusion XGBoost and defined four conditions: A = leaked structured + original text; B = leaked structured + clean text; C = clean structured (W24) + original text; D = clean structured (W24) + clean text. The total leakage effect is measured by $A-D$. The contribution of structured leakage is approximated by $B-D$, and the contribution of text leakage by $C-D$.

3 Results and Conclusion

Across all four tasks in TIMELY-Bench, performance inflation under leaked settings was driven almost entirely by future structured measurements rather than note content (Table 1). Total AUROC gains from leakage were +0.0154 (mortality), +0.0508 (prolonged ICU stay), +0.0463 (AKI progression), and +0.0399 (sepsis to septic shock). Modifying text to control for leakage yielded nearly identical performance, demonstrating that text leakage contributes negligible additional signal. This likely occurs because note-level embeddings pool and dilute isolated future-oriented sentences within broader clinical context. Consequently, rigorous control of structured lookahead must be a first-order design requirement for multimodal EHR research. While text leakage was minimal in this setting, it should not be generalised uncritically; future work will extend the benchmark to finer-grained (e.g., sentence- or span-level) text representations, broader note categories, and external multi-site validation.

Table 1: 2×2 leakage analysis across four tasks (early-fusion XGBoost, mean AUROC). A = leaked structured + original text; B = leaked structured + clean text; C = clean structured (W24) + original text; D = clean structured (W24) + clean text.

Task	**A**	**B**	**C**	**D**	$A-D$ **Total**	$B-D$ **Struct.**	$C-D$ **Text**
Mortality	0.9232	0.9231	0.9079	0.9079	+0.0154	+0.0153	+0.0000
Prolonged LOS	0.9368	0.9370	0.8856	0.8860	+0.0508	+0.0510	−0.0004
AKI progression	0.9176	0.9172	0.8709	0.8714	+0.0463	+0.0459	−0.0004
Sepsis to septic shock	0.9845	0.9844	0.9446	0.9446	+0.0399	+0.0399	+0.0000

4 Study Context

Ethics and approvals. This work uses de-identified MIMIC-IV data under PhysioNet credentialed access and data use agreements [6]. No direct patient contact or intervention was involved. Under King's College London research governance policy, secondary analysis of fully de-identified public datasets does not require separate ethical approval.

Funding. This work received no external funding.

Stakeholder involvement. No direct patient and public involvement was performed for this retrospective benchmarking study.

Data/method availability. Source data are available through credentialed PhysioNet access. Code and benchmark artefacts will be shared via the project repository and tagged release.

Conflicts of interest. The authors declare no competing interests.

References

[1] Khadanga S, Aggarwal K, Joty S, Srivastava J. Using Clinical Notes with Time Series Data for ICU Management. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019. p. 6432-7.

[2] Deznabi I, Iyyer M, Fiterau M. Predicting in-hospital mortality by combining clinical notes with time-series data. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021; 2021. p. 4026-31.

[3] Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop; 2019. p. 72-8.

[4] Johnson AEW, Pollard TJ, Mark RG. Reproducibility in critical care: a mortality prediction case study. In: Proceedings of the 2nd Machine Learning for Healthcare Conference. vol. 68 of Proceedings of Machine Learning Research; 2017. p. 361-76. Available from: https://proceedings.mlr.press/v68/johnson17a.html.

[5] Davis SE, Matheny ME, Balu S, Sendak MP. A Framework for Understanding Label Leakage in Machine Learning for Health Care. Journal of the American Medical Informatics Association. 2024;31(1):274-80.

[6] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data. 2023;10(1):1.

[7] Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. KDIGO Clinical Practice Guideline for Acute Kidney Injury. Kidney International Supplements. 2012;2(1):1-138.

[8] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):801-10.

[9] Styler WF IV, Bethard S, Finan S, Palmer M, Pradhan S, de Groen PC, et al. Temporal Annotation in the Clinical Domain. Transactions of the Association for Computational Linguistics. 2014;2:143-54.

Depression Severity Estimation via Speaker Diarization and Multi-Task Learning with Multimodal Cross-Attention

Tao Wang[1], Zhuoyuan Tang[1], Kai Yang[2], Li Yuan[3], Angus Roberts[1]

[1]Department of Biostatistics and Health Informatics, King's College London, UK.
[2]College of Economics, Shenzhen University, China.
[3]School of Software Engineering, South China University of Technology, China

Abstract

Background
Depression is a leading global health crisis affecting over 300 million individuals worldwide. Accurate and timely diagnosis is critical to enable effective treatment and reduce the broader burden. Current assessment practice relies on clinical interviews and standardised scales such as the PHQ-8, which are inherently subjective, resource-intensive, and susceptible to recall and rater bias. Recent advances in multimodal machine learning offer a promising path toward objective, scalable assessment by integrating acoustic, visual, and linguistic cues. However, existing methods typically process modalities through isolated pipelines [1] and predict depressive symptoms independently [2], failing to capture cross-modal interactions and inter-symptom dependencies, resulting in suboptimal feature fusion and incoherent severity estimates.

Methods
To address these limitations, this work presents a novel, end-to-end multimodal framework for depression detection. This framework includes three key components:

- **Speaker diarization**: Clinical interview recordings inherently contain both patient and interviewer speech. To prevent the model from learning spurious patterns driven by interviewer questioning strategy, we employ pyannote speaker diarization [3] to precisely segment audio and transcripts by speaker, strictly isolating patient utterances. This ensures that all downstream predictions are grounded exclusively in the patient's own linguistic and behavioural signals.
- **Multimodal cross-attention fusion:** Textual transcripts, facial action units, and acoustic features are first encoded using RoBERTa, OpenFace, and OpenSMILE respectively, before being passed through shared Bidirectional LSTM (BiLSTM) encoders to capture temporal dynamics within each modality. A multimodal cross-attention mechanism [4] is then applied to explicitly model the interactions across these three complementary data streams, enabling richer and more contextually informed feature fusion than isolated pipelines permit.
- **Multi-task severity prediction:** Rather than predicting the aggregate PHQ-8 score directly [1] or treating each questionnaire item independently [2], we formulate depression severity estimation as a Multi-Task Learning (MTL) problem in which each PHQ-8 item constitutes a distinct but interrelated task. This design allows the network to explicitly model inter-symptom dependencies, such as the relationship between sleep disturbance and fatigue, before producing a final severity estimate. The MTL framework is jointly optimised using a weighted Imbalanced Ordinal Log-Loss

(ImbOLL), which penalises misclassification of rare high-severity instances and directly addresses the pervasive class imbalance in clinical datasets.

We evaluate the proposed framework on the Extended Distress Analysis Interview Corpus (E-DAIC) [5], assessing predicted PHQ-8 severity scores against gold-standard labels using Concordance Correlation Coefficient, Mean Absolute Error, and Root Mean Squared Error, and benchmark performance against current state-of-the-art (SOTA) methods.

Results

The proposed framework was evaluated across five independent runs with different random seeds. It achieved an average CCC of 0.68 (higher is better), MAE of 3.62 (lower is better), and RMSE of 4.85 (lower is better), outperforming current SOTA results of CCC = 0.662, MAE = 3.95, and RMSE = 5.25, demonstrating the effectiveness of the proposed approach. Ablation studies confirm the contribution of each component. First, patient speech isolation via speaker diarization effectively removes interviewer noise, ensuring predictions are driven solely by patient-specific signals. Second, multimodal cross-attention fusion over RoBERTa, OpenFace, and OpenSMILE features captures complementary linguistic, facial, and prosodic cues that isolated pipelines fail to leverage. Third, the MTL formulation with cross-task attention and weighted ImbOLL produces coherent severity estimates while addressing class imbalance across PHQ-8 items. Furthermore, our analysis of PHQ-8 question dependency reveals that PHQ_1 (loss of interest), PHQ_2 (depressed mood), and PHQ_6 (feelings of failure) exhibit the strongest pairwise correlations ($r > 0.7$), directly validating our assumption that depressive symptoms are clinically interdependent and benefit from joint modelling.

Conclusion

This work demonstrates that explicitly modelling PHQ-8 item structure through MTL with multimodal cross-task attention, combined with patient speech isolation via speaker diarization, achieves state-of-the-art depression severity estimation. These results validate that jointly modelling inter-symptom dependencies and removing interviewer noise meaningfully improves automated assessment. Future work will focus on extending the framework to broader mental health conditions, improving model interpretability to support clinical trust, and prospective validation in real-world healthcare settings.

References

[1] Sadeghi M, Richer R, Egger B, Schindler-Gmelch L, Rupp LH, Rahimi F, et al. Harnessing multimodal approaches for depression detection using large language models and facial expressions. npj Mental Health Research. 2024;3(1):66.

[2] Mandal A, Atzil-Slonim D, Solorio T, Gurevych I. Enhancing Depression Detection via Question-wise Modality Fusion. In: Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025); 2025. p. 44-61.

[3] Bredin H, Yin R, Coria JM, Gelly G, Korshunov P, Lavechin M, et al. Pyannote. audio: neural building blocks for speaker diarization. In: ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE; 2020. p. 7124-8.

[4] Tsai YHH, Bai S, Liang PP, Kolter JZ, Morency LP, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th annual meeting of the association for computational linguistics; 2019. p. 6558-69.

[5] Ringeval F, Schuller B, Valstar M, Cummins N, Cowie R, Tavabi L, et al. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In: Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop; 2019. p. 3-12.

Comparison of Transformer Encoder-Based Text Classifiers in Identifying Canine Involvement in Road Traffic Accidents

Adam R. Williams[1], Peter-John M. Noble[1]

[1] University of Liverpool, Liverpool, United Kingdom

Introduction

Models based on the transformer encoder architecture (1) have seen significant performance increases since the release of BERT (2) in 2018. These improvements have been made both on general English language datasets through modifications to training task (RoBERTa (3), ELECTRA (4), deBERTa V3 (5)), new methods of positional encoding (MosaicBERT (6), deBERTa (7), ModernBERT (8)) and increases in context length (MosaicBERT (6), NomicBERT (9), ModernBERT (8)), and on domain specific datasets predominantly through domain adaptive pre-training (10). Performance improvements in Pre-Trained Language Models (PLM's) are measured on standard datasets such as MultiNLI (11), SQuAD (12), GLUE (13) and Super GLUE (14), while domain specific language model (DSLM) performance is typically measured using a downstream domain specific task (15-17). In this study, we aim to assess the performance of a variety PLM's and DSLM's on a domain specific task to examine whether improved in PLM performance is transferable to a task using domain specific language. The task used will be to fine-tune a classifier on each model to identify instances of canine involvement in road traffic accidents (RTA's). As a secondary result, we will identify risk factors associated with RTA involvement in dogs including breed, age, sex, neuter status and urban/rural location.

Methods and Data

This study was conducted using veterinary electronic health records collected through the Small Animal Veterinary Surveillance Network (SAVSENT) (18). SAVSNET is a veterinary research and surveillance system established in 2008 that collects real-time clinical data from a sentinel network of participating veterinary practices across the United Kingdom. The network automatically captures consultation-level data directly from integrated practice management systems; of most relevance here, these data include the clinical narrative entered by the attending veterinarian during appointments, and relevant demographic information about both the animal and its geographic location. Training data for the fine-tuning of classifiers was obtained by filtering all SAVSNET records with the species "dog" using a regular expression. A subset of records identified by regular expression were then manually classified into one of four categories (See Table 1), identifying any false positives that may have been captured by the regex. As the classes "RTA Involvement In Past" and "Possible RTA" were small compared to the "Confirmed RTA" and "Not RTA" classes, they were omitted from the dataset.

A list of models tested in this study can be seen in Table 1. Each model was fine-tuned using the tokenizer and optimizer defined in the pre-training of that model. The optimizers were initialized with the same hyperparameters (learning rate = $2e^{-5}$, Weight Decay=0.01, Batch Size=16) unless batch size needed reducing due to memory constraints (deBERTa V3-large (batch size=4), ModernBERT-large (batch size=12) and MosaicBERT (batch size=12)). For models designed to handle context windows wider than 512 tokens, the model was given a maximum sequence length equal to the number of tokens in the longest input sequence in the training dataset after tokenization using that models given tokenizer. This avoided excessive memory usage storing padding tokens while allowing the larger context window to be utilized. For models with a maximum context window length of 512 tokens, records with more than 512 tokens were truncated. Each model was fine-tuned using two NVIDIA RTX 4500 Ada Generation cards, with models being accessed via Huggingface and trained using the Huggingface Trainer Class. Early stopping was used to prevent overfitting, halting training when validation loss did not improve for three consecutive epochs. The loss function used was a standard cross entropy loss function.

After training, the best performing classifier was taken forward and used to classify all remaining records identified by regular expression as containing an RTA signal. All records identified as containing a positive signal were then used to train a multi-variable logistic regression model, calculating odds ratios for the variables of breed, age, sex, neuter status and location. Any records missing a date of birth, breed, sex, neuter status or location was removed from the dataset and where an animal had more than one record containing an RTA signal, one records was chosen at random to be taken forward to represent that animal.

Table 1. Class sizes and token length distributions of clinical records in RTA training dataset. All lengths were obtained after tokenizing using the BERT tokenizer

Class	Size Class	Mean Consult Length (WordPiece Tokens)	Median Consult Length (WordPiece Tokens)	Std Dev Consult Length (WordPiece Tokens)	Maximum Consult Length (WordPiece Tokens)
Confirmed RTA	238	161.12	127.00	132.15	1026
Not RTA	271	166.19	145.00	109.43	931
RTA involvement in past	19	82.95	57.00	64.11	294
Possible RTA	11	145.36	123.00	88.68	293

Table 2. PLM's and DSLM's on test in this experiment with release years

Pre-Trained Language Models	Domain Specific Language Models
BERT-base (2018) (2)	BioBERT (2020) (19)
BERT-large (2018) (2)	ClinicalBERT (2020) (17)
RoBERTa-base (2019) (3)	VetBERT (2020) (16)
RoBERTa-large (2019) (3)	PetBERT (2023) (15)
deBERTa V3-base (2021) (5)	DogBERT (2024)
deBERTa V3-large (2021) (5)	RoDogBERTa (2025)
MosaicBERT (2023) (6)	Clinical ModernBERT (2025) (20)
NomicBERT (2024) (9)	
ModernBERT (2025) (8)	

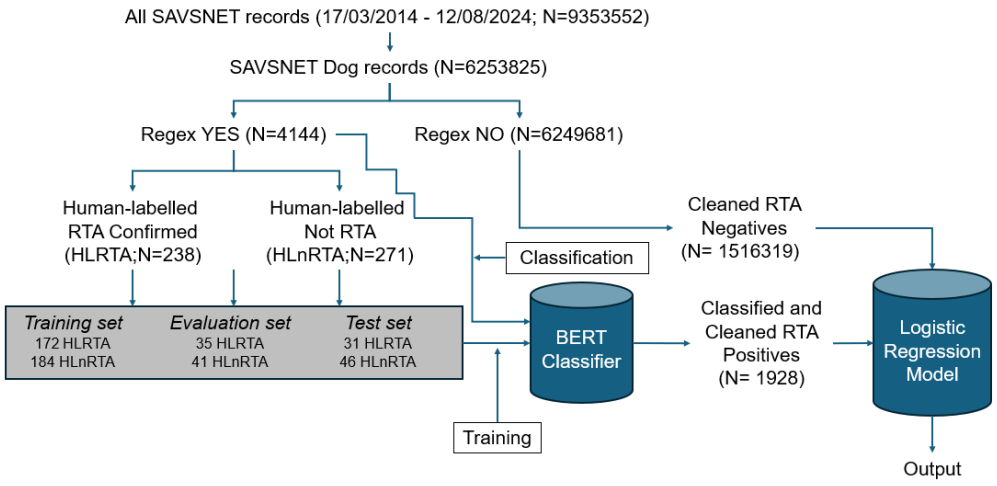


Figure 1. Data Flow Diagram for Classifying Instances of RTA positives

Results

Using F1-Score as the metric to judge model performance, the highest performing model was DogBERT (F1=89.4%) followed by NomicBERT (F1=88.5%), VetBERT (F1=87.1%), deBERTa V3-base (F1=86.2%) and both PetBERT and deBERTa V3-large (F1=85.7%). All results are detailed in Figure 2. In general, we see an improvement in PLM performance over time with models such as NomicBERT and deBERTa V3 showing improved performance over BERT-base, likely due to their architectural improvements. However, MosaicBERT, ModernBERT-base and ModernBERT-large showed poorer performance when compared to earlier BERT models iterations. We hypothesize that this is due to the nature of our data, with short consult lengths not allowing the model to make use of its wide context window. Of the DSLM's it appears that models with training corpora closer to the target corpora performed better, with DogBERT (domain adapted on canine specific veterinary text) performing best. However, two DLSM's exhibited poor performance. Clinical ModernBERT we hypothesize suffered from the same issues as ModernBERT. However, we expected improved performance from RoDogBERTa, given that it was domain adapted using a more model architecture. The fact that it achieved no performance gain over RoBERTa base (from which it was domain adapted) implies that a larger domain specific corpus is required to elicit performance improvements.

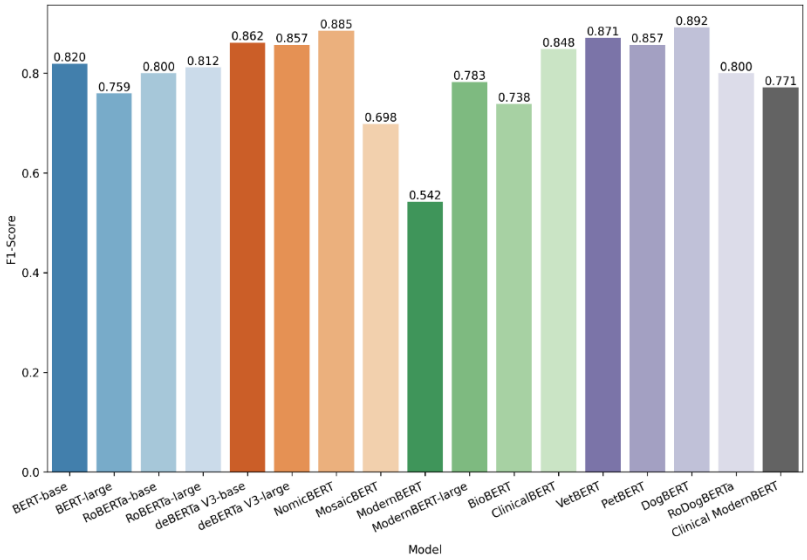


Figure 2. F1-Score of all PLM's and DSLM's at classifying RTA involvement of dogs from veterinary electronic health records

The results of the multi-variable logistic regression model can be seen in Figure 2. The model achieved an ROC AUC of 0.669, similar to the model using in the only other study these authors could find examining RTA instances in dogs (21). Our model suggests that that the most significant risk factor to a dog's likelihood of being involved in an RTA is age, with the odds ratio of a dog being involved in an RTA decreases as a dog ages (compared to a geriatric dog) (Young Adult (95% CI 3.16-5.12), Mature Adult (95% CI 2.43-3.8) and Senior (95% CI 1.27-2.05)). Age categories used are as defined by ND Harvey (22). Breed also appears to be a significant factor with the Generic Collie (95% CI 1.45-3.93), Hungarian Vizsla (95% CI 1.19-3.14) and Border Collie (95% CI 1.35-2.02) all showing increased odds ratios of RTA involvement compared with crossbreed. We hypothesize that this is likely due to behavioural factors such as chase behaviour which has been examined in other papers (23). The model also predicts that female dogs are less likely to be involved in RTA's than male dogs (95% CI 0.75-0.91) and that neutered dogs are less likely to be involved in RTA's than entire dogs (95% 0.72-0.89). However, these effects appear small when compared with breed and age. Location in an urban or rural environment does not appear to have a significant effect.

Table 3. Multivariable logistic regression model for risk factors associated with dogs involved in RTA's attending primary-care veterinary practices in the United Kingdom

Variable	Category	Odds Ratio	CI_95	p.value	Significance
Age Category at Consult	Puppy	0.85	0.63 - 1.14	0.2757	
	Juvenile	4.84	3.75 - 6.24	<0.001	***
	Young Adult	4.02	3.16 - 5.12	<0.001	***
	Mature Adult	3.04	2.43 - 3.8	<0.001	***
	Senior	1.61	1.27 - 2.05	<0.001	***
	Geriatric	Base			
Breed	Crossbreed	Base			
	Beagle	1.16	0.75 - 1.8	0.492	
	Bichon Frise	0.86	0.52 - 1.41	0.5407	
	Border Collie	1.65	1.35 - 2.02	<0.001	***
	Border Terrier	0.94	0.63 - 1.39	0.7416	
	Boxer	0.81	0.49 - 1.33	0.4037	
	Cavalier King Charles Spaniel	0.54	0.34 - 0.86	0.0094	**
	Chihuahua	0.68	0.48 - 0.96	0.0298	*
	Collie (Generic)	2.39	1.45 - 3.93	<0.001	***
	French Bulldog	0.48	0.33 - 0.69	<0.001	***
	German Shepherd Dog (Alsatian)	0.71	0.5 - 1.01	0.0556	
	Hungarian Vizsla	1.94	1.19 - 3.14	0.0073	**
	Jack Russell Terrier	1.01	0.81 - 1.26	0.9256	
	Pug	0.64	0.42 - 1	0.0482	*
	Retriever (Golden)	0.67	0.45 - 1.01	0.0566	
	Retriever (Labrador)	0.89	0.75 - 1.06	0.1921	
	Shih Tzu	0.66	0.47 - 0.92	0.0156	*
	Spaniel (Cocker)	1.01	0.84 - 1.23	0.8881	
	Spaniel (Springer)	0.82	0.6 - 1.12	0.2115	
	Staffordshire Bull Terrier	1.08	0.86 - 1.36	0.4882	
	West Highland White Terrier	0.57	0.36 - 0.91	0.0187	*
	Whippet	1.38	0.91 - 2.09	0.1337	
	Yorkshire Terrier	0.67	0.46 - 0.98	0.0413	*
Gender	Male				
	Female	0.82	0.75 - 0.91	<0.001	***
Neuter	No	Base			
	Yes	0.8	0.72 - 0.89	<0.001	***
Loctation	Urban	Base			
	Rural	0.93	0.84 - 1.04	0.199	

Conclusion

In this study we have shown that while optimum performance may still be achieved by a DSLM, domain adapted using an appropriate corpus, PLM's have closed the performance gap. However, careful consideration should be takin in model selection, matching the features of the model to that of the text to be classified. From this, we have been able to show that both age and breeds are significant contributing risk factors to RTA involvement, with juvenile, young adult and mature adult dogs being more at risk, along with the Generic and Border Collie and Hungarian Vizsla breeds. These results are consistent with the only other study found on this topic, with the exception of breed. This is likely down to our ability to obtain a larger sample size. We acknowledge that although we have taken steps to measure model performance and validate results, language models make errors. With best performing we expect roughly 10% of consults to be misclassified. To address this in future, we can attempt to expand our training dataset for future studies and address edge cases in training.

Study context

SAVSNET collates electronic health records from participating veterinary practices in near real-time with University of Liverpool ethics committee approval (RETH001081). This data is maintained within the University of Liverpool and not used in the prompting of any AI tool (e.g. ChatGPT, Claude, Gemini, Copilot) which may then use this information in its training of subsequent models. All models trained in this study were domain adapted and fine-tuned using University of Liverpool hardware and can be accessed through Huggingface. The dataset used in this study may not be made public due to the terms of our ethics agreement, but access to the SAVSNET database may be granted on reasonable request. Alternatively, a deidentified subset of the SAVSNET database (N=4415) may be accessed and used for proof of concept studies (https://www.liverpool.ac.uk/savsnet/research/using-savsnet-data-for-research/). This PhD project is jointly funded by the University of Liverpool, Engineering and Physical Sciences Research Council (EPSRC) and Dogs Trust. The authors declare no conflicts of interest.

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.
2. Devlin J, Chang M-W, Lee K, Toutanova K, editors. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers); 2019.
3. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692. 2019.
4. Clark K, Luong M-T, Le QV, Manning CD. Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:200310555. 2020.
5. He P, Gao J, Chen W. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:211109543. 2021.
6. Portes J, Trott A, Havens S, King D, Venigalla A, Nadeem M, et al. MosaicBERT: A bidirectional encoder optimized for fast pretraining. Advances in Neural Information Processing Systems. 2023;36:3106–30.
7. He P, Liu X, Gao J, Chen W. Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:200603654. 2020.
8. Warner B, Chaffin A, Clavié B, Weller O, Hallström O, Taghadouini S, et al., editors. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2025.
9. Nussbaum Z, Morris JX, Duderstadt B, Mulyar A. Nomic embed: Training a reproducible long context text embedder. arXiv preprint arXiv:240201613. 2024.
10. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al., editors. Don't stop pretraining: Adapt language models to domains and tasks. Proceedings of the 58th annual meeting of the association for computational linguistics; 2020.
11. Williams A, Nangia N, Bowman S, editors. A broad-coverage challenge corpus for sentence understanding through inference. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers); 2018.
12. Rajpurkar P, Zhang J, Lopyrev K, Liang P, editors. Squad: 100,000+ questions for machine comprehension of text. Proceedings of the 2016 conference on empirical methods in natural language processing; 2016.
13. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S, editors. GLUE: A multi-task benchmark and analysis platform for natural language understanding. Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP; 2018.
14. Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, et al. Superglue: A stickier benchmark for general-purpose language understanding systems. Advances in neural information processing systems. 2019;32.

15. Farrell S, Appleton C, Noble P-JM, Al Moubayed N. PetBERT: automated ICD-11 syndromic disease coding for outbreak detection in first opinion veterinary electronic health records. Scientific reports. 2023;13(1):18015.
16. Hur B, Baldwin T, Verspoor K, Hardefeldt L, Gilkerson J, editors. Domain adaptation and instance selection for disease syndrome classification over veterinary clinical notes. Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing; 2020.
17. Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:190405342. 2019.
18. Jones PH, Radford AD, Noble P-JM, Sánchez-Vizcaíno F, Menacere T, Heayns B, et al. SAVSNET: collating veterinary electronic health records for research and surveillance. Online Journal of Public Health Informatics. 2016;8(1):e61843.
19. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40.
20. Lee SA, Wu A, Chiang JN. Clinical modernbert: An efficient and long context encoder for biomedical text. arXiv preprint arXiv:250403964. 2025.
21. Harris GL, Brodbelt D, Church D, Humm K, McGreevy PD, Thomson PC, et al. Epidemiology, clinical management, and outcomes of dogs involved in road traffic accidents in the United Kingdom (2009–2014). Journal of veterinary emergency and critical care. 2018;28(2):140–8.
22. Harvey ND. How old is my dog? Identification of rational age groupings in pet dogs based upon normative age-linked processes. Frontiers in veterinary science. 2021;8:643085.
23. Cooper E, Zulch H, Mills DS. The role of breed versus personality and other demographic factors in predicting chasing behaviour in dogs. Applied Animal Behaviour Science. 2025;282:106463.

A Scalable Approach to Address the Lack of Labelled Clinical Free Text Data: Case Study for Venous Thromboembolism

Oscar Windrath-Carr, Maite Arribas
Imperial Clinical Analytics, Research and Evaluation (iCARE),
NIHR Imperial Biomedical Research Centre, Imperial College Healthcare NHS Trust &
Imperial College London, London, United Kingdom

Introduction

The development of labelled clinical free text data is essential for downstream tasks in digital health research, including training and evaluating natural language processing models. However, the availability of such data remains limited. Manual annotation of clinical documents is costly and time-consuming because it requires specialist expertise, and the presence of personal information in free text further restricts data access. Open-source datasets are also difficult to release due to the risk of breaching patient or staff confidentiality.

Several approaches have been proposed to address this data gap, including generating synthetic data, and weak supervision using structured signals derived from electronic health records (EHRs) to approximate labels (1). Many existing weak supervision methods rely on single structured signals, most commonly ICD-10 diagnosis codes. Although widely used for research and phenotyping, ICD-10 codes have well-documented limitations: coding accuracy varies across institutions due to their use in billing, and codes may reflect historical rather than incidence diagnoses, introducing misclassification and bias (2–4). Additionally, weak-supervision frameworks based on single heuristics may have limited precision (5–7).

This study addresses these limitations by evaluating a novel weak-supervision approach that combines multiple EHR-derived signals for more precise diagnosis labelling, offering an efficient alternative to manual annotation. We apply this framework to identify the presence of venous thromboembolism (VTE) from hospital admission data, including deep vein thrombosis (DVT) and pulmonary embolism (PE) (8). The performance of individual and combined signals is assessed against a clinician-annotated reference dataset.

Methods and Data

The manually labelled dataset included 492 acute hospital admissions from Imperial College Healthcare NHS Trust. Annotation was completed by two resident doctors using relevant radiology reports to label each admission as PE, DVT, Both or No VTE. Across 96 double-annotated encounters, raw agreement was 81.2%, corresponding to substantial inter-annotator agreement (Cohen's $\kappa = 0.68$, 95% CI 0.54–0.81). Disagreements were resolved by adjudication to produce the final labels. The final labelled, de-identified dataset contains 200 VTE cases including 123 labels for PE, 60 for DVT, 17 for Both and 292 cases with No VTE.

Weakly supervised labels were generated using structured EHR signals, including ICD-10 codes, radiology report types and prescription records. Positive labelling signals include: any relevant ICD-10 codes (9), presence of a relevant radiology report (US Doppler for DVT, CT pulmonary angiography (CTPA) for PE), first occurrence of a relevant ICD-10 code for a patient (to mitigate historical carry-over coding), a previously validated VTE free text ruleset positive for the relevant condition on a radiology report (10) and prescription of therapeutic anticoagulation. Signals were first evaluated individually. Performance was evaluated using precision, sensitivity and F1. 95% CIs were calculated using Wilson's method. Selected combinations were then constructed primarily to increase precision and reduce false positives whilst retaining sufficient sensitivity to produce a usable training dataset.

Negative labels were generated using a complementary strategy. Admissions without a relevant ICD-10 code but with a relevant radiology report containing negation patterns (e.g. "no pulmonary embolism" in CTPA reports; "no DVT" in US Doppler reports) were labelled as No VTE. Performance was evaluated against the reference dataset with 95% CIs.

Results

Figure 1 summarises the performance of individual and combined EHR signals for identifying positive VTE cases. Therapeutic anticoagulation exhibited the highest precision out of all individual signals (0.778) but the lowest sensitivity (0.420). The presence of a relevant radiology report alone demonstrated limited discriminatory value with low precision (0.486) and high sensitivity (0.850). Refining this rule with the rulesets applied to radiology reports improved precision modestly to 0.505 (+0.019) with moderate sensitivity (0.785). The presence of any relevant ICD-10 code produced high sensitivity (0.980) but low precision (0.573). Refining this rule to first-occurrence ICD-10 codes improved precision to 0.683 (+0.110) whilst maintaining relatively high sensitivity (0.895). The optimal balance between sensitivity and precision was achieved by combining the first occurrence of a VTE ICD-10 code for the patient with a VTE ruleset positive radiology report. This resulted in a precision of 0.882 and a sensitivity of 0.710 with an F1 score of 0.788.

The results for the negative labelling approach (target label, 'No VTE') showed a high precision of 0.983 (95% CI: 0.94-1.00), low sensitivity of 0.408 (0.35-0.47) and low overall F1 score of 0.576 (0.53-0.62).

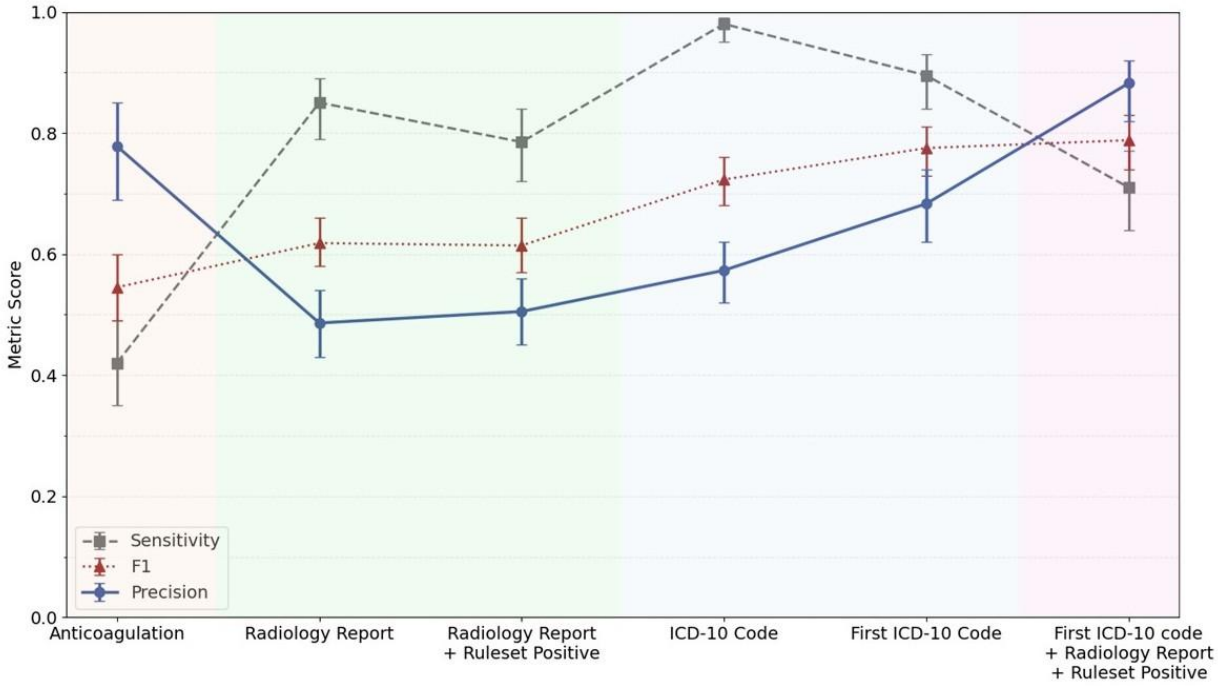


Figure 1. Performance of individual and combined rules for positive VTE labelling

Conclusion

Individual EHR-derived signals are insufficient for reliable weak supervision of VTE in isolation. ICD-10 codes alone demonstrate high sensitivity but inadequate precision, leading to substantial noise which presents a challenge for model development and validation. Requiring multiple EHR signals to positively label admissions for VTE presence allows for substantially improved precision whilst retaining acceptable sensitivity. This approach resulted in a high precision of 0.882 and a sensitivity of 0.710. The approach for negatively labelling radiology achieved a very high precision of 0.983 and a sensitivity of 0.408. Sensitivity was expected to be low due to the requirement of a relevant radiology procedure but owing to the predominance of No-VTE admissions, this approach produces a sufficient negatively labelled subset. Datasets constructed using this approach can enable the development of models for VTE risk prediction and automated case identification. Consequently, these models can support clinical decision-making by guiding preventive treatment to reduce avoidable in-hospital deaths. They can also facilitate service evaluations that inform policy and practice, such as refining the VTE risk assessment process for inpatients. Although demonstrated using VTE, this framework could be applied to other conditions with available ICD-10 codes and domain-specific document types, including imaging-based diagnoses (e.g. stroke), biomarker-defined acute events (e.g. myocardial infarction), or narrative-based psychiatric conditions (e.g. major depressive disorder).

Study context

This study was undertaken using EHR data from Imperial College Healthcare NHS Trust (ICHT), a large network of five hospitals providing acute and specialist care in North-West London, serving over 1.3 million patients annually. This data and research were enabled by the iCARE Digital Collaboration Space & Secure Data Environment (SDE) and used the iCARE team and data resources. The work was funded by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre (NIHR203323) with infrastructure support from the NIHR North-West London Patient Safety Research Collaboration (NIHR NWL PSRC, Ref. NIHR204292). This project has been reviewed and approved by the ICHT Data Protection Office and Caldicott Guardian. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

References

1. Alshaikhdeeb B, Hemedan AA, Ghosh S, Balaur I, Satagopam V. Generation of Synthetic Clinical Text: A Systematic Review [Internet]. arXiv; 2025 [cited 2026 Feb 18]. Available from: http://arxiv.org/abs/2507.18451 doi:10.48550/arXiv.2507.18451
2. Pendergrass SA, Crawford DC. Using Electronic Health Records to Generate Phenotypes for Research. Curr Protoc Hum Genet. 2019 Jan;100(1):e80. doi:10.1002/cphg.80 PubMed PMID: 30516347; PubMed Central PMCID: PMC6318047.
3. Horsky J, Drucker EA, Ramelson HZ. Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits. AMIA Annu Symp Proc. 2018 Apr 16;2017:912–20. PubMed PMID: 29854158; PubMed Central PMCID: PMC5977598.
4. Liu B, Hadzi-Tosev M, Eisa K, Liu Y, Lucier KJ, Garg A, et al. Accuracy of venous thromboembolism ICD-10 codes: A systematic review and meta-analysis. Thrombosis Update. 2024 Mar 1;14:100154. doi:10.1016/j.tru.2023.100154
5. Shen Z, Schutte D, Yi Y, Bompelli A, Yu F, Wang Y, et al. Classifying the lifestyle status for Alzheimer's disease from clinical notes using deep learning with weak supervision. BMC Med Inform Decis Mak. 2022 Jul 7;22(1):88. doi:10.1186/s12911-022-01819-4
6. Cusick M, Adekkanattu P, Campion TR, Sholle ET, Myers A, Banerjee S, et al. Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. Journal of Psychiatric Research. 2021 Apr 1;136:95–102. doi:10.1016/j.jpsychires.2021.01.052
7. Datta S, Roberts K. Weakly supervised spatial relation extraction from radiology reports. Jamia Open. 2023 Jul 1;6(2):ooad027. doi:10.1093/jamiaopen/ooad027
8. Overview | Venous thromboembolic diseases: diagnosis, management and thrombophilia testing | Guidance | NICE [Internet]. [cited 2026 Feb 12]. Available from: https://www.nice.org.uk/guidance/ng158
9. bmjopen-2015-008864supp_tables.pdf [Internet]. [cited 2026 Feb 12]. Available from: https://bmjopen.bmj.com/content/suppl/2015/11/11/bmjopen-2015-008864.DC1/bmjopen-2015-008864supp_tables.pdf
10. Deng J, Wu Y, Hayssen H, Englum B, Kankaria A, Mayorga-Carlin M, et al. Improving VTE Identification through Adaptive NLP Model Selection and Clinical Expert Rule-based Classifier from Radiology Reports [Internet]. arXiv; 2023 [cited 2026 Feb 12]. Available from: http://arxiv.org/abs/2309.12273 doi:10.48550/arXiv.2309.12273

PAIR-EHR: Transforming Clinical Case Reports into Structured EHR Representations

Chao Xu[1], Xiaolei Diao[1], Alec Diallo[2], Luo Mai[2], Yunsoo Kim[1,3], and Honghan Wu[1,3]

[1]University of Glasgow, Glasgow, UK
[2]University of Edinburgh, Edinburgh, UK
[3]University College London, London, UK

1 Introduction

Richly annotated electronic health record (EHR) data are essential for training and evaluating clinical natural language processing (NLP) systems, clinical decision support tools, and population health models [1, 2, 3, 4]. Access to such data is limited by privacy legislation, institutional governance requirements, and the substantial manual effort required for de-identification and annotation [5, 6]. Transforming freely available clinical narrative literature, such as case reports from PubMed Central (PMC)[7], into structured EHR representations provides an alternative approach. It enables the development of diverse and credible patient datasets at scale without the regulatory constraints associated with real data access [8, 9].

Recent advances in large language models (LLMs), such as GPT-5, have significantly improved the extraction of structured clinical information from unstructured text [10, 11, 12]. Despite these advancements, systematic comparative evaluations of these models for reconstructing complete EHR tables from discharge-style narratives are limited.

To address this gap, this study introduces **PAIR-EHR** as an automated two-stage pipeline that transforms PMC patient case reports into MIMIC-IV style EHR tables. The pipeline is evaluated across six LLMs on extraction breadth and semantic fidelity, using manual annotations as a reference standard.

2 Methods

2.1 Dataset and Annotation

Thirty PMC patient case reports were converted into discharge-style narrative summaries through a rule-based transformation, structured to mirror the format of MIMIC-IV discharge notes [5, 8] and covering presenting complaint, past medical history, medications, investigations, and discharge plan. The target output schema comprised seven MIMIC-IV `hosp`-module tables: `patients`, `admissions`, `diagnoses_icd`, `d_icd_diagnoses`, `prescriptions`, `labevents`, and `d_labitems`. Thirty notes were all selected for manual annotation of diagnoses, medications, and laboratory results by a clinical informatics expert.

Structured extraction can be applied directly to raw PMC case report text; however, a pilot study across 16 cases using two frontier LLMs (GPT-5.3 and Claude Sonnet 4.6) demonstrated that an optional rule-based transformation of case reports into discharge-style narrative summaries yields extraction volumes comparable to or marginally greater than those obtained from raw case reports across all seven tables. This format also promotes consistency with the distributional properties of real clinical discharge notes. On this basis, discharge-style summaries were adopted as input for the full-scale evaluation, which was expanded to 30 cases.

2.2 Models

Six LLMs were evaluated: GPT-5.3, Claude Sonnet 4.6, and DeepSeek-V3.2 as frontier general-purpose models; and MediPhi-Instruct, medgemma-4b-it, and KnowMedPhi3.5 as domain-specific biomedical models [10, 13, 14, 15, 16]. All models were evaluated in a zero-shot setting.

2.3 Prompting and Evaluation

The pipeline comprised two sequential prompting stages: structured clinical entity extraction from discharge notes into a constrained JSON format, followed by ICD10 code assignment for extracted diagnoses. JSON outputs were post-processed into relational CSVs conforming to the MIMIC-IV column schemas.

Semantic fidelity for the 30 annotated notes was assessed using micro-aggregated precision, recall, and F1 across three clinical domains: diagnoses, medications, and labs. A fuzzy matching scheme was applied to account for surface-form variation in clinical terminology, such as spelling variants and partial concept matches.

3 Results

	Diagnoses			**Medications**			**Labs**		
Model	P	R	F1	P	R	F1	P	R	F1
DeepSeek-V3.2	47.1	**74.4**	57.7	86.9	86.2	86.6	44.2	**51.7**	**47.7**
Claude Sonnet 4.6	53.1	70.5	60.6	85.3	**88.4**	86.8	43.2	50.3	46.5
GPT-5.3	**66.0**	62.2	**64.0**	**91.7**	**88.4**	**90.0**	57.5	18.6	28.1
MediPhi-Instruct	62.5	19.2	29.4	75.6	44.9	56.4	66.3	19.1	29.7
medgemma-4b-it	52.2	22.4	31.4	73.1	76.8	74.9	46.8	36.0	40.7
KnowMedPhi-3.5	56.3	31.4	40.3	80.2	52.9	63.8	**71.8**	24.0	36.0

Table 1: Semantic Fidelity Results. Precision (P), Recall (R), and F1 (%).

All six models generated outputs that conformed to the schema across the seven MIMIC-IV compatible tables. Frontier models extracted significantly more records from clinically substantive tables, including diagnoses and laboratory events, whereas domain-specific models produced considerably fewer entries. Semantic fidelity results (Table 1) indicate a consistent trend: frontier models achieved higher recall and overall F1 scores for diagnoses and medications, while domain-specific models demonstrated higher precision but markedly lower recall. Laboratory extraction was the most challenging task for all models. These findings indicate that structured output generation and instruction-following continue to be significant limitations for smaller domain-specific models.

4 Conclusion

PAIR-EHR presents a practical pipeline for generating MIMIC-IV–compatible EHR datasets derived from clinical literature. This approach demonstrates reproducibility across model families and establishes a benchmark for evaluating LLM-based clinical information extraction. Future research will expand the annotated evaluation set and integrate standardized laboratory term normalization.

5 Study context

Our study does not raise any ethical considerations as all the data used in the case study come from publicly available PubMed Central, and no patient-identifiable information was processed. Nonetheless, we acknowledge that the use of large language models for clinical information extraction introduces ethical considerations that warrant discussion. LLMs are non-deterministic and prone to confabulation. Such confabulations would constitute noise rather than signal in downstream applications. This limitation reinforces the need for manual clinical review of generated outputs and should be carefully considered before any downstream use of PAIR-EHR–generated datasets in clinical or research contexts. The authors kindly acknowledge funding from a PAIR (Population AI Research programme) EPSRC grant (UKRI2701).

References

[1] Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. Journal of the American Medical Informatics Association. 2018;25(5):530-7.

[2] Soysal E, Warner JL, Wang J, Jiang M, Harvey K, et al. Developing Customizable Cancer Information Extraction Modules for Pathology Reports Using CLAMP. Studies in health technology and informatics. 2019 Aug;264:1041-5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7359882/.

[3] Remy F, Demuynck K, Demeester T. Automatic Glossary of Clinical Terminology: a Large-Scale Dictionary of Biomedical Definitions Generated from Ontological Knowledge. In: Demner-fushman D, Ananiadou S, Cohen K, editors. The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. Toronto, Canada: Association for Computational Linguistics; 2023. p. 265-72. Available from: https://aclanthology.org/2023.bionlp-1.23.

[4] Johnson B, Bath T, Huang X, Lamm M, Earles A, et al.. Large language models for extracting histopathologic diagnoses from electronic health records. medRxiv; 2024. Pages: 2024.11.27.24318083. Available from: https://www.medrxiv.org/content/10.1101/2024.11.27.24318083v1.

[5] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data. 2023;10(1):1.

[6] Wu H, Wang M, Wu J, Francis F, Chang YH, Shavick A, et al. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. npj Digital Medicine. 2022;5(1):186.

[7] National Center for Biotechnology Information (NCBI). PubMed Central (PMC); 2000. Free full-text archive of biomedical and life sciences literature; accessed 2025. https://pmc.ncbi.nlm.nih.gov/.

[8] Zhao Z, Jin Q, Chen F, Peng T, Yu S. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. Scientific data. 2023;10 1:909. Available from: https://api.semanticscholar.org/CorpusID:266360591.

[9] Kweon S, Kim J, Kim J, Im S, Cho E, Bae S, et al. Publicly shareable clinical large language model built on synthetic clinical notes. In: Findings of the Association for Computational Linguistics: ACL 2024; 2024. p. 5148-68.

[10] OpenAI. GPT-5; 2025. Available from: https://openai.com/index/introducing-gpt-5/.

[11] Chen Z, Diao S, Wang B, Li G, Wan X. Are large language models ready for healthcare? A comparative study on clinical language understanding. arXiv preprint. 2023;arXiv:2304.05368.

[12] Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. npj Digital Medicine. 2023;6(1):195.

[13] DeepSeek-AI. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models; 2025. Available from: https://arxiv.org/abs/2512.02556.

[14] Anthropic. System Card: Claude Sonnet 4.6. Anthropic; 2026. Available from: https://www-cdn.anthropic.com/bbd8ef16d70b7a1665f14f306ee88b53f686aa75.pdf.

[15] Sellergren A, Kazemzadeh S, Jaroensri T, Kiraly A, Traverse M, Kohlberger T, et al. Medgemma technical report. arXiv preprint arXiv:250705201. 2025.

[16] Corbeil JP, Dada A, Attendu JM, Abacha AB, Sordoni A, Caccia L, et al. A Modular Approach for Clinical SLMs Driven by Synthetic Data with Pre-Instruction Tuning, Model Merging, and Clinical-Tasks Alignment. arXiv preprint arXiv:250510717. 2025.

Infusing Medical Hierarchies into Transformers

Yusuf Yildiz[1], Goran Nenadic[1] Meghna Jani[1], David A. Jenkins[1]
[1]University of Manchester, Manchester United Kingdom

1. Introduction

Transformer-based architectures, notably BEHRT(1), have demonstrated that longitudinal Electronic Health Record sequences can be effectively modelled to capture complex patient trajectories. However, these baseline models inherently treat clinical codes as flat, statistically independent tokens. This approach discards critical domain knowledge embedded in standard medical terminologies. When standard models treat these medically adjacent codes as separate tokens, they fail to leverage this intrinsic relationship and lose valuable contextual information. While infusing external medical ontologies into deep learning models is known to improve representation learning(2) how to achieve this integration within a Transformer architecture remains an active area of exploration. Our objective is to investigate how this integration can be practically realized through different mechanisms. In this study, we explore three approaches for infusing medical ontologies into a BERT-based framework for disease prediction. We implementing and compare input-level tokenization, embedding-level formulation, and external graph attention infusion to see impacts of the predictive precision.

2. Methodology

The core task is multi-label next-visit disease prediction at immediate, 6-month and 12-month intervals. The baseline model is a standard BEHRT architecture utilizing flat embeddings. To construct the ontology-aware variants, we extract hierarchical relationships from a standard medical ontology, defining a mapping from specific leaf concepts to their ancestral parent codes. We evaluate the following three approaches:

Approach 1: Ontology-Aware Tokenization (Input Level)

The simplest intervention alters the sequential input prior to embedding.

Approach 2: Hierarchical Embedding Formulation (Embedding Level)

This approach maintains the standard input sequence but alters the fundamental lookup tables within the model. We introduce a hierarchical tensor based on a pre-processed mapping of clinical concepts to their direct parents (code2parent). The modified embedding formulation for a given input token is defined as:

$$E_{Final} = E_{Code} + E_{Age} + E_{Segment} + E_{Position} + E_{Hierarcy}$$

Approach 3: Graph-Attention Knowledge Infusion (Attention Level)

Adapting the Graph-based Attention Model (GRAM)(3) framework, we represent the medical ontology as a DAG. Leaf nodes represent specific clinical concepts c_i observed in the EHR, while non-leaf nodes (c_a, c_c, c_g) represent broader ancestral concepts. The final representation g_i of a leaf concept is computed as a weighted sum of its basic embedding e_i and the embeddings of its ancestors via an attention mechanism, where attention weights govern the contextual contribution of each hierarchical level. These refined representations form an embedding matrix G

3. Experimental Setup & Evaluation Plan

Models are trained and evaluated using longitudinal patient data derived from the UK Biobank cohort(4). We define three distinct multi-label prediction tasks: predicting the immediate next clinical code, predicting incident codes within a 6-month window, and predicting incident codes within a 12-month window.

Predictive performance will be evaluated using both AUROC and the Average Precision Score (APS). While AUROC provides a macro-level view of discriminative ability, APS is ranking true positive incident conditions.

We plan to investigate the structural changes within the model's internal representations. By visualizing the learned embedding space aim to determine whether the infusion of hierarchical knowledge successfully clusters medically related, yet distinct, ICD-10 codes closer together compared to the isolated representations of the flat baseline.
Experiments are currently running. Comprehensive results will be provided in the final presentation.

4. Conclusion

Initial framework developments demonstrate that naive implementations of BERT on EHR data struggle with sparse, imbalanced clinical outcomes. By isolating the injection of ontological knowledge to three distinct architectural levels, this study will provide a comparative analysis of how structural medical knowledge influences the internal representations and downstream predictive precision of Transformer models.

Study Context

Ethics and Approvals: This study uses UK Biobank data. Requests to access these datasets should be directed to https://www.ukbiobank.ac.uk.
Funding: The author(s) declared that financial support was received for this work and/or its publication. Yusuf Yildiz was funded by the Republic of Türkiye Ministry of National Education. Meghna Jani is funded by a National Institute for Health and Care Research (NIHR) Advanced Fellowship [NIHR301413]. The views expressed in this publication are those of the authors and not necessarily those of the NIHR, NHS or the UK Department of Health and Social Care.
Data Availability: Due to patient privacy regulations, the raw EHR data cannot be made publicly available. Code repositories and synthetic sample data for the modified BEHRT and graph-attention mechanisms will be made available upon publication.
Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: Transformer for Electronic Health Records. Sci Rep. 2020 Apr 28;10(1):1. doi:10.1038/s41598-020-62922-y

2. Niu K, Lu Y, Peng X, Zeng J. Fusion of sequential visits and medical ontology for mortality prediction. J Biomed Inform. 2022 Mar 1;127:104012. doi:10.1016/j.jbi.2022.104012

3. Choi E, Bahadori MT, Song L, Stewart WF, Sun J. GRAM: Graph-based Attention Model for Healthcare Representation Learning [Internet]. arXiv; 2017 [cited 2025 Jun 15]. Available from: http://arxiv.org/abs/1611.07012 doi:10.48550/arXiv.1611.07012

4. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015 Mar;12(3):e1001779. doi:10.1371/journal.pmed.1001779 PubMed PMID: 25826379; PubMed Central PMCID: PMC4380465.

Exploring limitations of guideline-grounded Clinical Decision Support Systems by comparison with clinical practice

Agathe Zecevic[1,2], Angus Roberts[3], and Sebastian S. Zeki[1,4]

[1]King's College London, Faculty of Life Sciences & Medicine, London, UK
[2]Guy's and St Thomas' NHS Foundation Trust, Clinical Scientific Computing, London, UK
[3]King's College London, Biostatistics and Health Informatics, Institute of Psychiatry, Psychology & Neuroscience, London, UK
[4]Guy's and St Thomas' NHS Foundation Trust, Gastroenterology department, London, UK

1 Introduction

Surveillance for many premalignant conditions is guideline-driven, which has encouraged NLP clinical decision support systems (CDSS) to embed guidelines as the primary decision logic, improving their transparency and auditability [1]. However, such systems often assume that real-world care is primarily guideline-executing and that guideline-relevant evidence is consistently documented in the clinical narrative. In practice, adherence to guidelines is frequently impacted by external factors and documentation practices evolve over time [2, 3, 4]. It is important to understand how these factors impact CDSS.

First, follow-up timing can be driven by competing indications and alternative pathways or by procedural and patient-specific factors that are not listed in guidelines [5]. Second, national guidelines evolve over time, potentially impacting the robustness of CDSS. In Barrett's oesophagus (BO), the 2014 British Society of Gastroenterology (BSG) update promoted more standardised minimum dataset reporting, including segment length using Prague criteria [6, 7]. Such documentation drift can induce a temporal dataset shift that may affect NLP extraction robustness.

The primary aim of this study is to classify the reasons for discrepancies between current clinical practice and guideline-grounded CDSS. The secondary aim is to characterise documentation drift between guideline updates. Finally, we test how drift affects CDSS performance.

2 Methods and Data

We conducted three complementary experiments using retrospective evaluations of endoscopy reports for patients with two premalignant conditions that undergo guideline-based [6, 8] endo-

scopic surveillance, Barrett's Oesophagus (BO) and Colorectal Polyps (CP). In both domains, the CDSS combines NLP extraction from endoscopy free text with deterministic guideline logic to compute a recommended surveillance interval and follow-up date. We compare CDSS outputs with observed booked follow-up dates from routine care.

(1) Clinical deviation analysis

We performed a secondary mixed-methods review of discrepant cases between AI pipelines and observed care. A discrepancy was defined as an absolute difference of ≥ 6 months between the CDSS-recommended follow-up date (index date + recommended interval) and the booked follow-up date. For the BO cohort, a subset (N = 106) of discrepancies from the original CDSS retrospective analysis was randomly sampled. For the CP cohort all the discrepancies from the retrospective study (N = 41, two exclusions) were included. The discrepancies were classified as potentially unintentional, intentional or alternative pathway management according to the following definitions:

- *Unintentional*: No documented rationale, which can reflect booking/administrative issues, documentation gaps or clinical error.
- *Intentional*: Clinically justified divergence with the rationale documented in the endoscopy report or the booking
- *Alternative pathway management*: follow-up primarily driven by a different pathway or additional clinical findings.

The cases were reviewed by a single expert with additional clinician adjudication for ambiguous cases.

(2) Documentation drift

We compared BO endoscopy reports from **pre-2014** (2012–2013) versus **post-2014** (2015–2016) (balanced $n = 751$ per dataset). Drift was characterised using descriptive statistics (length, tokens), key-field presence (Prague criteria), TF–IDF classifier-based two-sample test (C2ST) [9, 10] with repeated 5-fold cross-validation and embedding-based drift analysis [11]. We additionally re-ran the drift tests with confounder masking: we removed Prague score tokens and restricted the analysis to endoscopists appearing in both time periods to check whether separability persisted.

(3) Impact on real-world CDSS

To isolate the impact of documentation drift on downstream performance, we evaluated one component of a post-2014-trained CDSS on 200 randomly sampled pre-2014 BO reports using the same post-2014 definitions. The model is tasked to classify Barrett's length used downstream for guideline logic: {No, Short, Long, Insufficient}. We report per-class precision/recall/F_1 (baseline post-2014 results from the EndominerAI BO study [12]).

3 Results

Deviation analysis

Table 1 summarises the results of the discrepancy analysis. Across both domains, the most

frequent discrepancy category was alternative pathway, indicating cases where guideline surveillance is not the actual decision policy.

Table 1: Deviation taxonomy among analysed discrepant cases (counts and % within clinical domain).

Domain	$n_{analysed}$	**Unintentional**	**Intentional**	**Alternative pathway**
BO	106	39 (36.8)	22 (20.8)	45 (42.5)
CP	41	12 (29.3)	12 (29.3)	17 (41.5)

Alternative pathways discrepancies reflected heterogeneous real-world drivers (symptom-led reassessment, cancer pathways, other clinical conditions management). Intentional deviations showed recurring clinical patterns such as failed/incomplete procedures or heightened risk context.

Documentation drift. Post-2014 reports were longer and Prague criteria mentions increased from 39.0% to 60.7%, consistent with the 2014 BSG emphasis on standardised BO length reporting [6, 7]. The TF–IDF C2ST strongly discriminated pre- vs post-2014 reports (Accuracy = 0.78 ± 0.02, ROC AUC 0.87 ± 0.02, mean across repeated 5-fold CV) and separability persisted after Prague masking (Accuracy = 0.78 ± 0.02, ROC AUC 0.88 ± 0.02) and within endoscopist-overlap subsets (Accuracy = 0.74 ± 0.04, ROC AUC 0.84 ± 0.03), suggesting broader lexical changes.

Impact on real-world CDSS Table 2 summarises the post-2014 CDSS performance on pre-2014 endoscopy reports. Despite measurable documentation drift evidenced by the previous analysis, the CDSS performance was largely preserved, with modest recall degradation for some classes (Long: $1.00 \rightarrow 0.93$; No Barrett's: $0.94 \rightarrow 0.89$), indicating robustness to documentation shift while still highlighting the need to monitor model performance over time.

4 Conclusion

Real world clinical decisions rely on a number of factors that cannot always be captured in guidelines and guideline-based CDSS. This underscores the importance of partnering CDSS with clinicians both at the development and output stage as well as incorporating broader data sources. Our study also suggests CDSS monitoring should be event-driven around guideline updates, combining drift detection with targeted re-evaluation of downstream performance.

5 Study context

All the experiments in the study were performed under ethics approval of the GERRI board at Guy's and St Thomas' NHS Foundation Trust. Data cannot be shared publicly due to patient confidentiality and governance restrictions. Code is available upon request. The authors declare no competing interests.

Table 2: Temporal robustness of a CDSS.
Post-2014 performance is taken from [12] (GSTT external test set; $N = 100$). Pre-2014 performance is from this study ($N = 200$), annotated under the same variable definitions.

	Post-2014 (baseline)			Pre-2014 (this study)		
Category	Precision	Recall	F_1	Precision	Recall	F_1
No Barrett's	0.96	0.94	0.95	0.96	0.89	0.93
Short	1.00	0.96	0.98	0.95	0.99	0.97
Long	1.00	1.00	1.00	1.00	0.93	0.96
Insufficient	0.73	0.85	0.79	0.83	0.94	0.88

References

[1] Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. npj Digital Medicine. 2020;3:17.

[2] Roumans CAM, et al. Adherence to recommendations of Barrett's esophagus surveillance guidelines: a systematic review and meta-analysis. Endoscopy. 2020;52(1).

[3] Butler DM, et al. Adherence to Post-polypectomy Surveillance Guidelines at a Large District General Hospital. Cureus. 2023;15(2):e35516.

[4] Holmberg D, et al. Adherence to clinical guidelines for Barrett's esophagus. Scandinavian Journal of Gastroenterology. 2019.

[5] Agency for Healthcare Research and Quality. Taxonomy of Override Reasons for Patient-Centered Clinical Decision Support (PC CDS) Recommendations; 2024. NCBI Bookshelf report.

[6] Fitzgerald RC, di Pietro M, Ragunath K, et al. British Society of Gastroenterology guidelines on the diagnosis and management of Barrett's oesophagus. Gut. 2014;63(1):7-42.

[7] Sharma P, Dent J, Armstrong D, et al. The development and validation of an endoscopic grading system for Barrett's esophagus: the Prague C & M criteria. Gastroenterology. 2006;131(5):1392-9.

[8] Rutter MD, East J, Rees CJ, et al. British Society of Gastroenterology/Association of Coloproctology of Great Britain and Ireland/Public Health England post-polypectomy and post-colorectal cancer resection surveillance guidelines. Gut. 2020;69(2):201-23.

[9] Lopez-Paz D, Oquab M. Revisiting Classifier Two-Sample Tests. arXiv. 2016.

[10] Koch LM, Baumgartner CF, Berens P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. npj Digital Medicine. 2024 May;7(1). Available from: `http://dx.doi.org/10.1038/s41746-024-01085-w`.

[11] Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Association for Computational Linguistics; 2019. .

[12] Zecevic A, Jackson L, Zhang X, Pavlidis P, Dunn J, Trudgill N, et al. Automated decision making in Barrett's oesophagus: development and deployment of a natural language processing tool. npj Digital Medicine. 2024;7(1):312.

A Comparison Study of Three Pipelines for Barrett's Oesophagus Surveillance Prediction

Xinyue Zhang[1], Agathe Zecevic[2], Sebastian Zeki[2], and Angus Roberts[1]

[1]King's College London, London, United Kingdom
[2]Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom

1 Introduction

In Barrett's oesophagus (BO) surveillance, endoscopy and pathology reports contain key information required to determine surveillance intervals according to British Society of Gastroenterology (BSG) guidelines[1]. Recent work has explored two main paradigms for automated surveillance prediction from clinical reports: 1) Structured extraction pipelines, which extract entities and relations and then apply rule-based decision logic; 2) Report classification models, which directly assign report labels without explicit intermediate structure[2].

In this work we compare three approaches on performance, explainability, annotation requirements, and computational efficiency: 1) Extraction-based pipeline using eREBEL[3]; 2) LLM-based structured extraction[3, 4]; 3) Report-classification models (EndoBERT and PathBERT)[2]

2 Methods and Data

2.1 Datasets and Task

Models were evaluated on two UK hospital datasets: GSTT[1] and KCH[2]. These datasets contain annotated endoscopy and pathology reports with report-level labels and surveillance outcomes. The task is to classify reports into clinically relevant categories such as[3, 2]: 1) Endoscopy: Long, Short, NoBarretts, Insufficient; 2) Pathology: DysplasiaOrCancer, IM, No_IM, Insufficient. These report labels are then can be mapped deterministically to surveillance intervals using a rule-based algorithm derived from BSG guidelines.

2.2 Model Paradigms

Figure 1 illustrates the three paradigms considered.

Extraction-based pipeline produces extractions such as: Prague scores, Barrett's length, intestinal metaplasia. These structured outputs provide interpretable evidence for clinical decisions.

- **eREBEL-based[3]** models perform joint entity and relation extraction from reports and feeds the extracted structured information into a rule-based algorithm for surveillance decision prediction.
- **LLM-based[3, 4]**, such as Phi-4 (14B) can also produce structured outputs from clinical narratives through prompting, producing JSON-like outputs of the same clinical entities used by rule-based algorithms.

Report classification models[2], such as EndoBERT and PathBERT, directly assign report-level labels without extracting intermediate entities. These models rely on contextual representation learning within the transformer encoder to capture clinically relevant signals.

[1]https://www.guysandstthomas.nhs.uk/
[2]https://www.kch.nhs.uk/

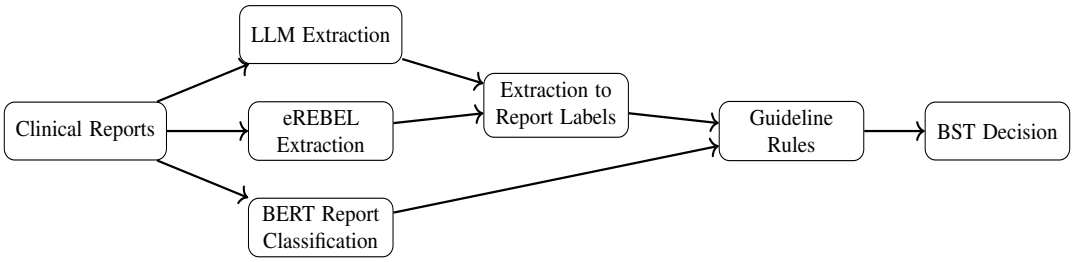


Figure 1: Three modelling paradigms for Barrett's surveillance prediction.

3 Results

Dataset / Task	Metric	Phi-4 (14B)	Qwen-2.5 (14B)	DeepSeek Qwen-2.5 (14B)	eREBEL (0.4B)	Endo/PathBERT (0.1B)
Pathology (GSTT)	Weighted avg	**0.97** (0.93, 0.99)	0.96 (0.93, 0.99)	0.92 (0.87, 0.96)	0.71 (0.64, 0.78)	0.92
	Inference Time	28.82	30.81	70.47	2.03	0.03
Endoscopy (GSTT)	Weighted avg	0.92 (0.87, 0.97)	0.94 (0.89, 0.97)	**0.95** (0.91, 0.99)	0.83 (0.77, 0.89)	**0.95**
	Inference Time	28.82	30.81	70.47	2.03	0.03
Pathology (KCH)	Weighted avg	**0.92** (0.87, 0.95)	0.89 (0.83, 0.93)	0.90 (0.85, 0.95)	0.80 (0.74, 0.87)	0.88
	Inference Time	27.64	28.26	66.23	2.07	0.03
Endoscopy (KCH)	Weighted avg	0.82 (0.76, 0.87)	0.84 (0.78, 0.89)	**0.87** (0.82, 0.92)	0.75 (0.68, 0.82)	**0.87**
	Inference Time	27.64	28.26	66.23	2.07	0.03

Table 1: Comparison of weighted-average F1 scores and inference time

Table 1 compares endoscopy and pathology classification performance between the general-purpose LLMs and two fine-tuned domain-specific models. Overall, Phi-4 and Endo/PathBERT show consistently strong performance across tasks, whereas the extraction models (Phi-4 and eREBEL extraction) achieves produce structured and interpretable outputs.

However, LLMs incur substantially higher inference time and computational cost than eREBEL and BERT-based models. In contrast, fine-tuned models require additional annotation and retraining when adapting to new tasks or datasets, whereas LLM-based systems can often be adapted through prompt modifications alone. Extraction-based approaches also provide reusable structured outputs that can support downstream queries and related tasks. Table 2 summarises these trade-offs across key criteria.

Model	Endo/PathoBERT	eREBEL	LLM
Performance	★★★	★	★★★
Explainability	★	★★★	★★★
Deployment	★★★	★★	★
Space Efficiency	★★★	★★	★
Development	★	★	★★★
Reusability	★	★★	★★★

Table 2: Comparison of three pipelines across evaluation criteria.

4 Conclusion

We compared three clinical NLP paradigms for predicting BST-related report labels from endoscopy and pathology reports. Our results show that LLMs and specialised BERT classifiers achieve similarly strong report-level performance, while extraction-based models provide more interpretable structured outputs. These approaches involve different trade-offs between predictive performance, explainability, computational cost, and ease of adaptation. Such trade-offs are important when selecting models for clinical decision support systems operating under practical deployment constraints.

5 Study context

This study was approved by the institutional review board (IRAS ID: 257283). It is funded by the King's Centre for Doctoral Training in Data-Driven Health, supported by EPSRC funding. High-performance computing resources are provided by King's CREATE Trusted Research Environment. Data is accessible to passported researchers upon request. Project code will be made publicly available. This work was conducted in collaboration with Guy's and St Thomas' NHS Foundation Trust (GSTT) and King's College Hospital (KCH).

References

[1] Fitzgerald RC, Di Pietro M, Ragunath K, Ang Y, Kang JY, Watson P, et al. British Society of Gastroenterology guidelines on the diagnosis and management of Barrett's oesophagus. Gut. 2014;63(1):7-42.

[2] Zecevic A, Jackson L, Zhang X, Pavlidis P, Dunn J, Trudgill N, et al. Automated decision making in Barrett's oesophagus: development and deployment of a natural language processing tool. NPJ Digital Medicine. 2024;7(1):312.

[3] Zhang X. Automating Surveillance Timing for Barrett's Oesophagus: An Information Extraction Approach Using Natural Language Processing [Doctoral thesis]. King's College London; 2026. Available from: https://kclpure.kcl.ac.uk/portal/en/studentTheses/automating-surveillance-timing-for-barretts-oesophagus.

[4] Zhang X, Zecevic A, Zeki S, Roberts A. Improving Barrett's Oesophagus Surveillance Scheduling with Large Language Models: A Structured Extraction Approach. In: Proceedings of the 24th Workshop on Biomedical Language Processing; 2025. p. 176-89.