Dora explores Clinically Relevant Information in EHRs using NER

Martin Sundahl Laursen
Department of Clinical Biochemistry
Odense University Hospital
martin.sundahl.laursen@rsyd.dk

Lina Elkjær Pedersen
Department of Clinical Biochemistry
Odense University Hospital

Josefine Bak H Adelhelm
Department of Clinical Biochemistry
Odense University Hospital

Rasmus Bank Lynggaard
Department of Clinical Biochemistry
Odense University Hospital

Pernille Just Vinholt
Department of Clinical Biochemistry, Odense University Hospital
Department of Clinical Research, University of Southern Denmark

Abstract

Retrieving relevant information from unstructured electronic health records is time-consuming and prone to error, reducing time available for direct patient care. We present Dora, a Danish clinical named entity recognition model that builds on prior work by Laursen et al. (2023a). Dora identifies six types of clinical entities to support medical information retrieval: diseases, symptoms/findings, diagnostics, treatments, anatomies, and results. The model achieves an exact boundary macro F1 score of 0.922 and overlap boundary score of 0.945. A prospective clinical utility evaluation shows that Dora reliably extracts relevant information for physicians. A bias analysis indicates slightly reduced performance on psychiatric notes, with minimal overall differences.

1 Introduction

Health care professionals, particularly medical doctors (MDs), need to retrieve information from electronic health records (EHRs) regarding diagnoses, symptoms, medications, treatments, etc. This process is time-consuming, carries the risk of overlooking important information, and ultimately reduces the time available for direct patient care (Laursen et al., 2023b). Furthermore, the health care data is in an unstructured format in the EHR. The EHR system may include a basic "find on page" function, which allows users to search for specific words or phrases within the visible text. However, this method is vulnerable to inaccuracies such as misspellings, abbreviations, and typographical errors.

Previous work has shown that natural language processing methods, particularly Named Entity Recognition (NER) models, can effectively identify clinical entities in EHR text (Jiang et al., 2011; Alsentzer et al., 2019).

Notable results in English clinical NER are Stanza (Qi et al., 2020; Zhang et al., 2021) with micro F1 0.881 and BioBERT (Lee et al., 2020) with micro F1 0.867 on identifying problems, tests, and treatments in the i2b2 dataset (Uzuner et al., 2011). In Scandinavian clinical NER, Laursen et al. (2023a) achieved an entity-level macro F1 of 0.601 for exact boundary matching and 0.682 for overlap boundaries when identifying diseases, symptoms (including abnormal findings), diagnostics, treatments, anatomies, and results in Danish EHR text. In Swedish, RoBERTa Large by AI Sweden achieved a token-level micro F1 of 0.779 when compared to eight other encoder models on identifying diagnoses, findings, body parts, and drugs in the Stockholm EPR Clinical Entity Corpus (Vakili et al., 2025; Skeppstedt et al., 2014).

Few medical machine learning studies, however, extend beyond reporting internal test set performance and do not assess real-world clinical impact and utility (Kelly et al., 2019; Ghassemi et al., 2020; Rajpurkar et al., 2022).

In this paper, we extend the Danish Clinical NER model by Laursen et al. (2023a), retraining it on an expanded and re-annotated dataset with updates to the annotation scheme, preprocessing, postprocessing, and model training.

We present a new Danish clinical NER model, Dora, that presents substantial improvements in model performance. We demonstrate the clinical utility in a prospective real-world evaluation and evaluate bias.

2 Methods

In this section, we first describe the data sources for the model's development and evaluation cohort. We then outline the model's development

and present the different evaluations conducted to assess the model.

2.1 Data Sources

We used data from two different EHR systems of Odense University Hospital in the Region of Southern Denmark for development and evaluation. The COSMIC cohort consisted of EHRs from the COSMIC system (Cambio, CGI, Denmark) from November 2015 to September 2020. The EPJ cohort consisted of all EHRs from the EPJ SYD (Systematic, Denmark) system from February 2022 to November 2023.

2.2 Model Development

The development of the first iteration of the NER model was previously described by Laursen et al. (2023a). Here, we focus on refinements made to the annotation scheme, dataset, system architecture, and model development.

2.2.1 Annotation

We built on the clinical event annotation scheme proposed by Laursen et al. (2023a), with one key revision to improve usability for healthcare professionals: the Symptom entity, which used to include symptoms and pathological findings, now includes symptoms and all clinical findings—either normal or pathological.

The dataset from Laursen et al. (2023a) was re-annotated by a MD to reflect the revised scheme and iteratively extended with paragraphs from the COSMIC and EPJ cohorts using active learning and a locally developed annotation tool. Targeted data augmentation was applied to address specific errors.

2.2.2 Dataset

Our dataset contained 158,839 total entities, almost triple the size of the original dataset, split into training, validation and test sets. Splits were stratified to maintain a balanced distribution across entity labels, see Table 1.

2.2.3 System Architecture

We adapted the Princeton University Relation Extraction system (PURE) (Zhong and Chen, 2021), using code with minor modifications from Laursen et al. (2023a).

PURE classifies entities from constructed span-embeddings, which is the concatenation of the contextual embeddings of the start and end tokens, along with a learned span-width embedding (Zhong and Chen, 2021). For full architectural details, we refer to the original work. Our modifications applied to the implementation by Laursen et al. (2023a) include:

- **Preprocessing:** Lowercasing, removing non-printable/control characters, converting HTML entities to Unicode, and mapping uncommon accented or special characters to standard equivalents.
- **Postprocessing:** After prediction, overlapping spans with the same label are merged. When overlaps have different labels, a voting mechanism selects the most likely label.

2.2.4 Development

We followed Laursen et al. (2023a) in extracting contextual embeddings using a Danish clinical ELECTRA encoder (Pedersen et al., 2022; Clark et al., 2020). Spans ranged from 1–10 tokens. Each of the start, end, and width components had size 256.

We trained using AdamW (Loshchilov and Hutter, 2019) (weight decay 0.001, batch size 32), early stopping (patience 6), and learning rate scheduling (patience 3, factor 0.2). No class weighting was applied.

Optimal learning rates (search space in parenthesis) were 5e-5 (5e-6–7.5e-5) for the encoder and 5e-4 (5e-5–7.5e-4) for the classifier. The optimal span classifier configuration was one (0–2) 1024-unit (256–1024; plateaued at 1024) hidden layer with ReLU activation and 0.3 dropout.

Model selection was based on the span-level macro F1 score on the validation set, excluding the negative class. We report the best model's entity-level recall, precision, and F1 score on the test set, using exact and overlapping boundary matching (Chinchor and Sundheim, 1993). We report a confusion matrix based on overlap matching, which better reflects clinical utility due to the often ambiguous boundaries of clinical entities.

2.3 Clinical Evaluation

The aim of the clinical evaluation was to assess the model's clinical utility and potential bias on an evaluation cohort stratified by gender (male/female), age group (child/adult/senior), and 15 diagnoses. To ensure diverse diagnoses, two MDs selected five diagnoses within each medical area; medical, psychiatry, surgical, from The Danish Med-

	Train (% of row total)	Validation (% of row total)	Test (% of row total)	TOTAL (% of column total)
Paragraphs	18,001 (80%)	2,206 (10%)	2,258 (10%)	22,465 (100%)
		Clinical events		
Disease	7,198 (81%)	821 (9%)	887 (10%)	8,906 (6%)
Symptom	37,692 (80%)	4,467 (10%)	4,808 (10%)	46,967 (30%)
Treatment	22,218 (80%)	2,806 (10%)	2,774 (10%)	27,798 (18%)
Diagnostic	21,631 (80%)	2,782 (10%)	2,654 (10%)	27,067 (17%)
Anatomy	25,444 (80%)	3,104 (10%)	3,234 (10%)	31,782 (20%)
Result	13,024 (80%)	1,714 (11%)	1,581 (10%)	16,319 (10%)
TOTAL	127,207 (80%)	15,694 (10%)	15,938 (10%)	158,839 (100%)

Table 1: Distribution of clinical event types in the training, validation, and test sets.

ical Classification System (SKS) (Danish Health Data Authority, n.d.), which is based on the International Classification of Diseases 10th revision (ICD-10) (World Health Organization, 2016). Diagnoses spanning all genders and age groups and with a high likelihood of mentions of varied clinical entities like symptoms, diagnostics, and treatments were chosen:

- **Medical:** asthma, diabetic ketoacidosis (type 1), epilepsy, pneumonia, rheumatoid arthritis
- **Psychiatry:** autism, depression, eating disorder, generalised anxiety disorder, suicide attempt/self-harm
- **Surgical:** appendicitis, hernia, ileus, epistaxis, tibia fracture

We then randomly sampled EHRs from the EPJ cohort that included either a ICD-10 code or a textual mention of one of these 15 diagnoses.

2.3.1 Clinical Utility Evaluation

To evaluate the model's clinical utility, a MD manually reviewed its output on the evaluation cohort. For each EHR, the model's extracted entities were shown in a spreadsheet containing one row per entity with its label and context window. The full EHR text was provided for reference. EHRs were included iteratively, seeking a stratified sample of three different EHRs for each combination of diagnosis, gender, and age (n=270).

The MD assessed whether the model output included at least one mention of: 1) the disease entity for the target diagnosis, 2) symptoms, 3) diagnostic tool, and 4) treatment relevant for that diagnosis. If the diagnosis was missing from predictions, the full EHR was reviewed to confirm its presence. If the diagnosis was absent, the EHR was not included. When any expected entity was missing, the full EHR was checked to determine if the model had failed to extract it.

257 samples were included. The cohort consisted of 132 females and 125 males, including 85 children, 86 adults, and 86 seniors. Two groups were entirely absent: female children with depression and senior males with eating disorder.

We calculated the detection rate per entity label.

The 15 diagnoses and expected clinical findings for each entity are presented in Appendix B.

2.3.2 Bias analysis

We conducted a structured bias analysis across gender, age group, and medical area.

From the evaluation cohort, we sampled three random patient EHRs (>5 notes available) per combination (n=270). Each patient was represented by four random notes (>49 characters per note to avoid minimal or templated content) (n=1,080).

Model predictions were corrected by a MD to establish ground truth. Entity-level F1 scores were calculated per patient, with micro and macro averages across labels (Chinchor and Sundheim, 1993). We report summary statistics for entity counts by medical area.

To ensure robust metrics given the short text span per patient, we applied conservative rules when averaging to handle missing entities:

- **No ground truths, some predictions:** recall excluded; precision and F1 set to 0.
- **Some ground truths, no predictions:** precision excluded; recall and F1 set to 0.
- **No ground truths or predictions:** all metrics excluded.

We further bootstrapped with 9,999 resamples per individual variable (i.e., each gender, age group,

and medical area) to produce 95% confidence intervals (CIs) by entity label and micro and macro average, using these to assess systematic model bias (Steyerberg et al., 2001).

3 Results

This section presents the results of the evaluation of the model performance, clinical utility, and potential biases.

3.1 Test Set Performance

The model achieved F1 scores above 0.90 across all entity types and evaluations. Macro F1 was 0.922 with exact boundary matching and 0.945 with overlap. Ignoring labels, the detection macro F1 with overlap reached 0.962. Detailed results are shown in Table 2.

	TEST SET					
	Exact boundary			Overlap boundary		
	F1	Prec	Recall	F1	Prec	Recall
Disease	0.914	0.921	0.906	0.936	0.939	0.932
Symptom	0.902	0.917	0.888	0.930	0.943	0.918
Treatment	0.926	0.932	0.920	0.953	0.957	0.949
Diagnostic	0.941	0.943	0.938	0.957	0.958	0.956
Anatomy	0.940	0.950	0.930	0.968	0.974	0.962
Result	0.907	0.910	0.905	0.930	0.931	0.929
Micro avg	0.922	0.930	0.913	0.946	0.953	0.940
Macro avg	0.922	0.929	0.915	0.945	0.950	0.941
Detection	0.932	0.941	0.924	0.962	0.969	0.956

Table 2: Model performance metrics on the test set. Prec = precision; Avg = average; Detection = Matching the text span regardless of the assigned label.

Figure 1 shows the confusion matrix for overlapping boundary matching. 3.0% of model detections were spurious, while 4.4% of ground truth spans were not detected. Of all spurious classifications, 36.8% were symptoms. The model missed 6.2% of symptoms and 5.6% of results.

3.2 Clinical Utility Evaluation

The model identified the diagnosis and at least one relevant symptom in all 257 patients (100% detection). Relevant treatments were detected in 99.2% of patients, missing only two cases: epilepsy ("at se an"—wait and see) and hernia ("reponere"—reposition/reduction). Diagnostic procedures were identified in 99.6% of cases, with one autism case missing "ADOS" and "WISC" assessment tools.

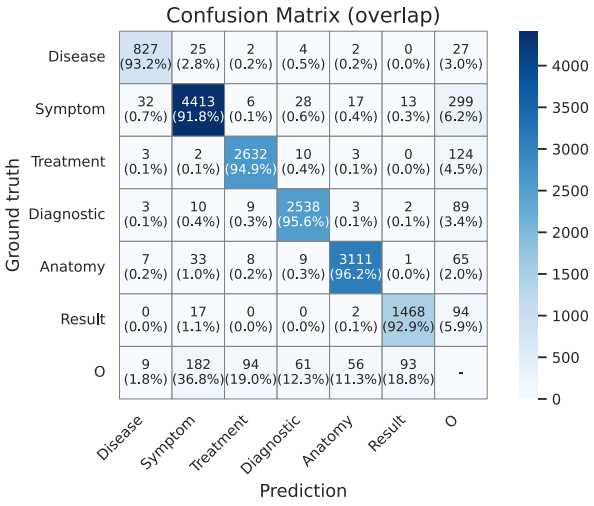

Figure 1: Confusion matrix for the model on the test set with overlapping boundary matching. O = Non-entity spans.

3.3 Bias Evaluation

Appendix Table A1 shows mean, median and range of entity counts by medical area. Psychiatric notes mention more symptoms on average (36) than medical (23) and surgical notes (20), with a wider range (0–237 vs. 0–90 and 0–86, respectively). They include fewer anatomies (7 vs. 11 and 12, respectively) and results (6 vs. 14 and 11, respectively).

Figure 2 shows the bootstrapped 95% CIs for macro and micro averaged F1 scores for comparison inside groups. The CI for children is non-overlapping and lower than for seniors but overlap with adults. The psychiatry CI is non-overlapping and lower than the medical and surgical CIs. The observed differences in means for the non-overlaps are at or below 0.017. All other CIs overlap.

Figure 3 shows the bootstrapped 95% CIs for F1 scores for each entity by group. The CIs for diagnostic, anatomy, and result entities overlap inside all groups. In contrast, for disease and symptom entities, the CIs for psychiatry are non-overlapping and lower than those for medical and surgical. For treatment entities, the psychiatry CI is lower than medical but overlap with surgical. The observed differences in means for the non-overlaps are at or below 0.033.

Mean F1 scores for all group levels range from 0.958 to 1.000, with full details on CIs reported in Appendix Table A2 and A3.

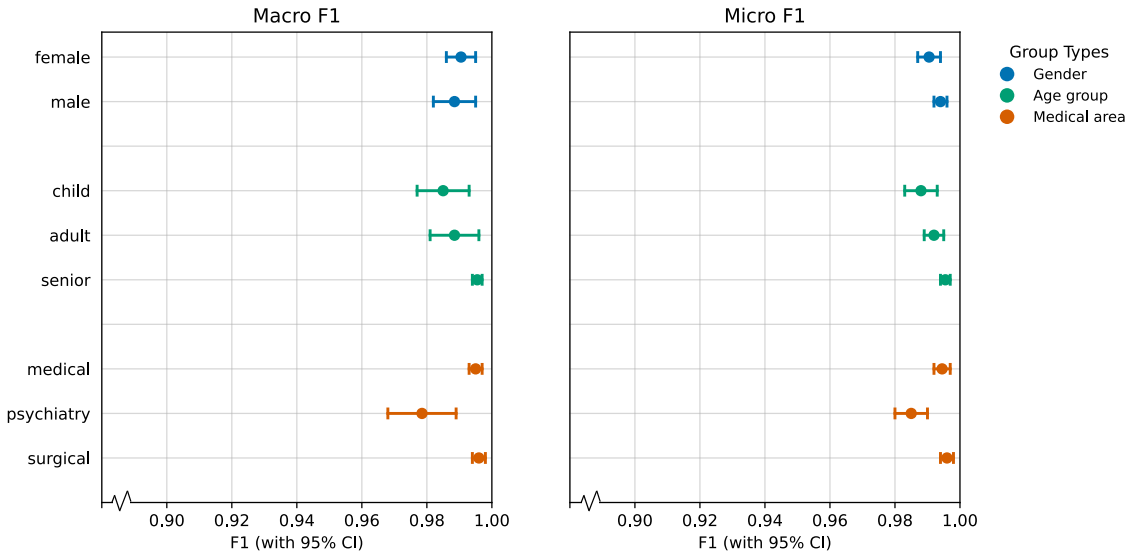

Figure 2: Bootstrapped 95% confidence intervals (CIs) for macro (left) and micro (right) averaged F1 scores by group. Note that comparison is only possible between the levels of each group, not between groups.

4 Discussion

We saw a substantial improvement in model performance, with exact boundary macro F1 increasing from 0.601 in the original work to 0.922, and overlap boundary F1 from 0.682 to 0.945. Likely causes include more consistent annotation by a single MD rather than six in the original work (Laursen et al., 2023a), a tripled dataset size, improved postprocessing with span merging and voting, and an updated annotation scheme that includes both symptoms, and normal and pathological findings under the Symptom category—reflecting their often similar context in clinical text and simplifying the classification task.

The excellent performance of the prospective utility evaluation of the model shows how it can be used to retrieve all relevant information for physicians managing patients of all ages and genders for the diagnoses included in the evaluation. Given the heterogeneous clinical presentations of the 15 diagnoses evaluated, these results suggest promising potential for broader implementation across all ICD-10 diagnoses.

The findings from our bias study indicate a small but consistent reduction in model performance on psychiatric notes, with minimal effects observed in other groups. While statistical significance was not formally assessed, these differences likely stem from the distinct structure and content of psychiatric notes. Based on clinical experience, psychiatric notes tend to be longer. They also mention more symptoms and contain fewer references to anatomy, results, and diagnostic tests (see Appendix Table A1). These factors suggest that the model could benefit from additional psychiatric notes in training data, although the current performance differences remain very small.

5 Conclusion

We present Dora, a Danish clinical NER model that identifies key clinical entities: diseases, symptoms (including normal and pathological findings), diagnostics, treatments, anatomies, and results. Dora achieves substantial improvements over the original model, with a macro F1 score of 0.922 for exact boundary matching and 0.945 for overlapping boundaries. Prospective utility evaluation demonstrates excellent performance in extracting relevant information for physicians. Our bias study reveals a small but consistent performance reduction on psychiatric notes, with minimal variation in other groups, though overall differences remain very small.

Limitations

While the bias analysis offers valuable insights, several limitations remain; firstly, using four notes per patient may not fully represent the medical condition in case of complex or chronic illnesses. To address this, we applied conservative metrics and bootstrapping in order to improve robustness. Secondly, ground truth labels for the bias study were

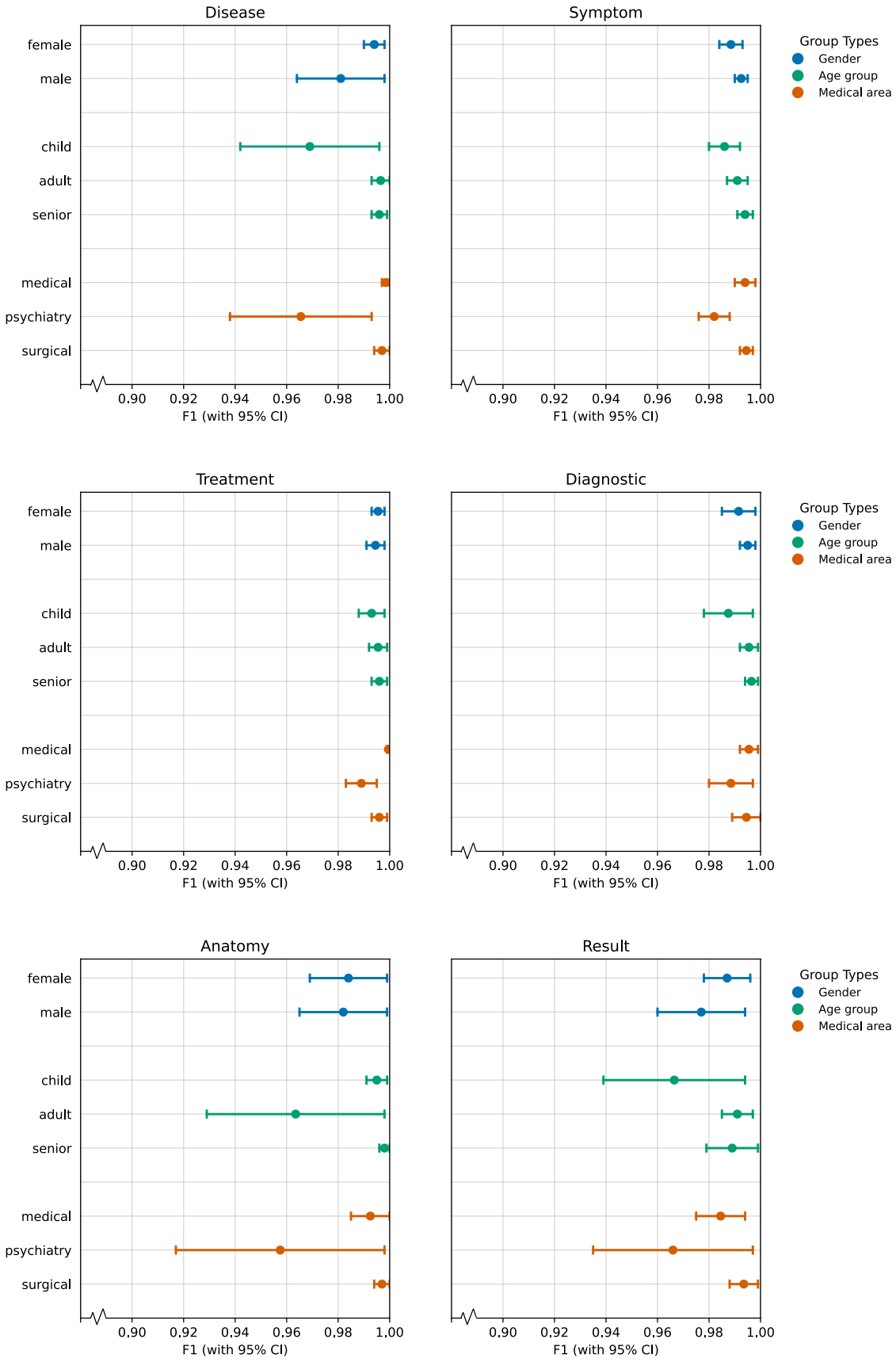

Figure 3: Bootstrapped 95% confidence intervals (CIs) for F1 scores for each entity by group. Note that comparison is only possible between the levels of each group, not between groups.

created by correcting model predictions, which, while efficient, may have influenced the annotations. Finally, notes were randomly sampled from a 1.5-year period during which the patient was given the relevant diagnosis. This approach ensures a diverse range of note types per medical area, improving generalisability. It may, however, introduce noise as some notes risk not being strongly representative of their originating medical area.

We cannot rule out that some individual sentences from the evaluation cohort may also appear in the training data. However, since evaluation was performed on full EHRs, the presence of single duplicate sentences, which are common due to standard phrasing in EHRs, is unlikely to impact results.

Ethics Statement

This study was conducted using clinical data accessed with appropriate institutional permissions. All data usage complied with relevant ethical guidelines and data protection regulations, and was approved by the data providers.

The dataset and model are not publicly available due to sensitive content. Please contact us for sharing options.

References

AI Sweden. RoBERTa Large. Accessed: 2025-07-30.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, WA Redmond, and Matthew BA McDermott. 2019. Publicly available clinical bert embeddings. *NAACL HLT 2019*, page 72.

Nancy Chinchor and Beth M Sundheim. 1993. Muc-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Danish Health Data Authority. n.d. The danish medical coding classification system. Accessed: 2025-06-05.

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. 2020. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191.

Min Jiang, Yukun Chen, Mei Liu, S Trent Rosenbloom, Subramani Mani, Joshua C Denny, and Hua Xu. 2011. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5):601–606.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17:1–9.

Martin Laursen, Jannik Pedersen, Rasmus Hansen, Thiusius Rajeeth Savarimuthu, and Pernille Vinholt. 2023a. Danish clinical named entity recognition and relation extraction. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 655–666.

Martin S Laursen, Jannik S Pedersen, Rasmus S Hansen, Thiusius R Savarimuthu, Rasmus B Lynggaard, and Pernille J Vinholt. 2023b. Doctors identify hemorrhage better during chart review when assisted by artificial intelligence. *Applied clinical informatics*, 14(04):743–751.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jannik S Pedersen, Martin S Laursen, Cristina Soguero-Ruiz, Thiusius R Savarimuthu, Rasmus Søgaard Hansen, and Pernille J Vinholt. 2022. Domain over size: Clinical electra surpasses general bert for bleeding site classification in the free text of electronic health records. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. Ai in health and medicine. *Nature medicine*, 28(1):31–38.

Maria Skeppstedt, Maria Kvist, Gunnar H Nilsson, and Hercules Dalianis. 2014. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of biomedical informatics*, 49:148–158.

Ewout W Steyerberg, Frank E Harrell Jr, Gerard JJM Borsboom, MJC Eijkemans, Yvonne Vergouwe, and

J Dik F Habbema. 2001. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of clinical epidemiology*, 54(8):774–781.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Thomas Vakili, Martin Hansson, and Aron Henriksson. 2025. Sweclineval: A benchmark for swedish clinical natural language processing. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 767–775.

World Health Organization. 2016. *International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10)*. https://icd.who.int/browse10/2016/en.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical english model packages for the stanza python nlp library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61.

A Bias Evaluation: Detailed Results

		Mean	Median	Range
Disease				
	medical	4.66	3	(0 - 27)
	psychiatry	5.82	3	(0 - 41)
	surgical	3.57	3	(0 - 18)
Symptom				
	medical	22.57	18	(0 - 90)
	psychiatry	35.98	19	(0 - 237)
	surgical	19.77	17	(0 - 86)
Treatment				
	medical	12.96	10	(0 - 45)
	psychiatry	13.04	10	(0 - 87)
	surgical	15.24	12	(0 - 50)
Diagnostic				
	medical	21.16	19	(1 - 78)
	psychiatry	12.10	8	(0 - 83)
	surgical	16.52	15	(0 - 50)
Anatomy				
	medical	11.02	7.5	(0 - 46)
	psychiatry	6.96	3	(0 - 47)
	surgical	12.06	9	(0 - 71)
Result				
	medical	13.72	12	(0 - 58)
	psychiatry	6.02	3	(0 - 31)
	surgical	11.04	8	(0 - 39)

Table A1: Mean, median and range of entity counts by medical area for the bias evaluation.

	Macro F1	Micro F1
Gender		
female	0.991 (0.986 - 0.995)	0.991 (0.987 - 0.994)
male	0.989 (0.982 - 0.995)	0.994 (0.992 - 0.996)
Age group		
child	0.985 (0.977 - 0.993)	0.988 (0.983 - 0.993)
adult	0.989 (0.981 - 0.996)	0.992 (0.989 - 0.995)
senior	0.996 (0.994 - 0.997)	0.996 (0.994 - 0.997)
Medical area		
medical	0.995 (0.993 - 0.997)	0.995 (0.994 - 0.997)
psychiatry	0.979 (0.968 - 0.989)	0.985 (0.980 - 0.990)
surgical	0.996 (0.994 - 0.998)	0.996 (0.994 - 0.998)

Table A2: Bootstrapped macro and micro F1 scores with 95% confidence intervals reported by group levels for the bias evaluation.

	Disease	Symptom	Treatment
Gender			
female	0.994 (0.990 - 0.998)	0.989 (0.984 - 0.993)	0.996 (0.993 - 0.998)
male	0.981 (0.964 - 0.998)	0.993 (0.990 - 0.995)	0.995 (0.991 - 0.998)
Age group			
child	0.969 (0.942 - 0.996)	0.986 (0.980 - 0.992)	0.993 (0.988 - 0.998)
adult	0.997 (0.993 - 1.000)	0.991 (0.987 - 0.995)	0.996 (0.992 - 0.999)
senior	0.996 (0.993 - 0.999)	0.994 (0.991 - 0.997)	0.996 (0.993 - 0.999)
Medical area			
medical	0.999 (0.997 - 1.000)	0.994 (0.990 - 0.998)	1.000 (0.999 - 1.000)
psychiatry	0.966 (0.938 - 0.993)	0.982 (0.976 - 0.988)	0.989 (0.983 - 0.995)
surgical	0.997 (0.994 - 1.000)	0.995 (0.992 - 0.997)	0.996 (0.993 - 0.999)

	Diagnostic	Anatomy	Result
Gender			
female	0.992 (0.985 - 0.998)	0.984 (0.969 - 0.999)	0.987 (0.978 - 0.996)
male	0.995 (0.992 - 0.998)	0.982 (0.965 - 0.999)	0.977 (0.960 - 0.994)
Age group			
child	0.988 (0.978 - 0.997)	0.995 (0.991 - 0.999)	0.967 (0.939 - 0.994)
adult	0.996 (0.992 - 0.999)	0.964 (0.929 - 0.998)	0.991 (0.985 - 0.997)
senior	0.997 (0.994 - 0.999)	0.998 (0.996 - 1.000)	0.989 (0.979 - 0.999)
Medical area			
medical	0.996 (0.992 - 0.999)	0.993 (0.985 - 1.000)	0.985 (0.975 - 0.994)
psychiatry	0.989 (0.980 - 0.997)	0.958 (0.917 - 0.998)	0.966 (0.935 - 0.997)
surgical	0.995 (0.989 - 1.000)	0.997 (0.994 - 1.000)	0.994 (0.988 - 0.999)

Table A3: Bootstrapped F1 scores and 95% confidence intervals by entity type and group levels.

B Clinical Utility Evaluation: Expected Findings

MEDICAL			
Disease	**Symptom**	**Diagnostic**	**Treatment**
Epilepsia [DG40]	Seizures Impaired consciousness Tongue bite Urination Amnesia	Blood samples Imaging Electro- encephalogram	Antiseizure medicine
Asthma [DJ45]	Dyspnoea Cough	Pulse Oximetry Imaging Blood samples a-puncture pH Pulmonary function test	Bronchodilator Oxygen Steroid
Diabetic ketoacidosis type 1 [DE101]	Polyuria/polydipsia Respiratory changes Nausea/vomiting Foetor ex ore Abdominal pain Weakness/fatigue Impaired consciousness	Blood samples a-puncture Glucose Urine sample	Insulin Fluid therapy
Rheumatoid arthritis [DM05, DM08]	Pain Swelling Redness Heat of joint(s) Fever Fatigue Other systemic symptoms	Blood samples Imaging	Anti-inflammatory drugs Immunomodulatory drugs Analgesics
Pneumonia [DJ189]	Dyspnoea Cough Fever	Pulse Oximetry Imaging Blood samples a-puncture pH	Bronchodilator Oxygen Steroid Antibiotics Fluid therapy

Table B1: Expected clinical findings in the health record for each medical diagnosis by entity type.

PSYCHIATRIC			
Disease	**Symptom**	**Diagnostic**	**Treatment**
Generalized anxiety [DF411]	Anxiety Headache Restlessness Pain Tension Fear Sleep disturbances Autonomic hyperactivity Tension	Psychiatric assessment	Psychotherapy Antidepressants CNS depressants
Depression [DF33]	Poor concentration Feelings of excessive guilt or low self-worth Hopelessness Thoughts about dying or suicide Disrupted sleep Changes in appetite or weight Feeling very tired or low in energy	MDI or Hamilton scale Psychiatric assessment	Antidepressants Psychotherapy Sleep medication
Autism [DF840]	Deficits within: Communication, interaction, and behaviour	Psychiatric assessment	Psychotherapy Sleep medication
Suicide attempt /self-injury [DZ915A]	Intentional cause of injury on oneself	Psychiatric assessment	Psychotherapy Antipsychotics CNS depressants
Eating disorder [DF50]	Disturbance in one's eating behaviors that affect the person's physical or mental health	Psychiatric assessment BMI	Psychotherapy Enteral/parenteral nutrition therapy

Table B2: Expected clinical findings in the health record for each psychiatric diagnosis by entity type.

SURGICAL			
Disease	**Symptom**	**Diagnostic**	**Treatment**
Appendicitis [DK35, DK37, DK379]	Pain Fever Nausea/vomiting	Abdominal examination Blood samples Imaging	Surgery/appendectomy Antibiotics Analgesics
Epistaxis [DR040C, DR040A, DR040B]	Bleeding from nose nose or mouth	Blood samples Rhino endoscopy Imaging	Compressive therapy Ablation Haemostatics Transfusion Fluid therapy
Fracture of tibia [DS821, DS823]	Pain Swelling Loss of function Displacement	Examination Imaging	Analgesics Fixation Surgery Antibiotics
Hernia [DK409]	Pain Protrusion Fever Nausea/vomiting	Abdominal examination Imaging Blood samples	Surgery Antibiotics Analgesics
Ilieus [DK567]	Pain Fever Nausea/vomiting	Abdominal examination Imaging Blood samples	Surgery Antibiotics Analgesics

Table B3: Expected clinical findings in the health record for each surgical diagnosis by entity type.