Adverse Event Extraction from Discharge Summaries: A New Dataset, Annotation Scheme, and Initial Findings

Imane Guellil[1], **Salomé Andres**[1], **Atul Anand**[1], **Bruce Guthrie**[1],
Huayu Zhang[1], **Abul Hasan**[2], **Honghan Wu**[3,2], **Beatrice Alex**[1],
[1]University of Edinburgh, [2]University College London, [3]University of Glasgow,
Correspondence: imane.guellil.ig@gmail.com

Abstract

In this work, we present a manually annotated corpus for Adverse Event (AE) extraction from discharge summaries of elderly patients, a population often underrepresented in clinical NLP resources. The dataset includes 14 clinically significant AEs—such as falls, delirium, and intracranial haemorrhage, along with contextual attributes like negation, diagnosis type, and in-hospital occurrence. Uniquely, the annotation schema supports both discontinuous and overlapping entities, addressing challenges rarely tackled in prior work. We evaluate multiple models using FlairNLP across three annotation granularities: fine-grained, coarse-grained, and coarse-grained with negation. While transformer-based models (e.g., BERT-cased) achieve strong performance on document-level coarse-grained extraction (F1 = 0.943), performance drops notably for fine-grained entity-level tasks (e.g., F1 = 0.675), particularly for rare events and complex attributes. These results demonstrate that despite high-level scores, significant challenges remain in detecting underrepresented AEs and capturing nuanced clinical language. Developed within a Trusted Research Environment (TRE), the dataset is available upon request via DataLoch and serves as a robust benchmark for evaluating AE extraction methods and supporting future cross-dataset generalisation.

1 Introduction

The automatic identification and extraction of Adverse Events (AEs) from clinical texts is a critical area of research in Natural Language Processing (NLP), with important applications in pharmacovigilance and patient safety. AEs refer to adverse outcomes caused by medical interventions, and their accurate detection is essential for supporting clinical decision-making and improving healthcare monitoring systems.

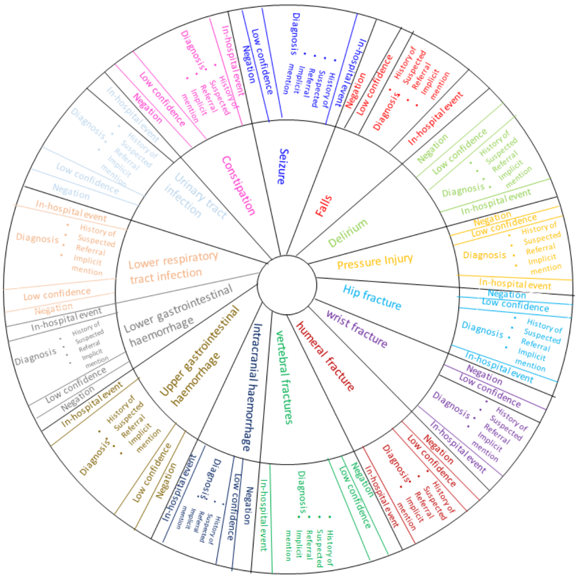


Figure 1: AE entities and attributes

Among clinical documents, discharge summaries-particularly those concerning elderly patients-pose unique challenges due to the complexity, variability, and ambiguity of their narrative structure (Murphy et al., 2023).

While recent advances in machine learning, particularly transformer-based models, have improved AE extraction, model performance remains heavily dependent on the availability of high-quality annotated corpora (Mahendran and McInnes, 2021). In practice, however, such datasets are limited, especially those tailored to clinical narratives, and existing resources often fail to address key linguistic challenges such as discontinuous and overlapping named entities. These phenomena are especially prevalent in clinical texts, where the structure of medical language can result in multi-part, noncontiguous mentions or nested concepts. For instance, the sentence "The wrist X-ray performed in A&E confirms the presence of a fracture" includes a discontinuous mention ("wrist" and "fracture")

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28532–28562
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

referring to a single AE. Similarly, overlapping entities arise when a shared span belongs to more than one annotation, as in "urinary and faecal incontinence", where "incontinence" is shared between two distinct AE labels. Such complex cases challenge standard Named Entity Recognition (NER) models, which were originally developed for flat entity structures and are often ill-equipped to handle biomedical texts that demand more expressive annotation. As the field evolves toward unified NER—encompassing flat, overlapping, and discontinuous structures—clinical NLP must adopt more sophisticated annotation schemas and evaluation strategies to reflect real-world use cases. To address these limitations, we present a manually annotated corpus of discharge summaries from elderly patients, in which AEs are systematically identified and labeled using a clinically informed annotation guideline. Our corpus includes 14 clinically significant AEs (e.g., falls, delirium, hip fractures, urinary tract infections, intracranial haemorrhages), each annotated with four contextual attributes: negation, in-hospital occurrence, diagnosis type (history_of, suspected, referral, implicit_mention), and low confidence. The annotation supports both fine-grained and coarse-grained extraction, and explicitly incorporates discontinuous and overlapping entity spans. The complete entity-attribute schema is shown in Figure 1.

To benchmark this dataset, we evaluate multiple pre-trained language models using the FlairNLP framework, including BioBERT and variants of BERT, across multiple annotation granularities. Our goal is to create a practical and challenging evaluation framework that mirrors the real-world conditions and annotation challenges encountered in clinical EHR data.. Our dataset is available upon request via DataLoch, and offers a valuable foundation for future work in AE extraction, dataset generalization, and unified NER in clinical NLP.

2 Related work

The extraction of adverse events (AEs) using NLP has become a crucial area of research due to its potential to improve pharmacovigilance and enhance patient safety (Guellil et al., 2024b). Much of the research in this domain can be categorised into three primary areas: AE classification (Tafti et al., 2017; Ellenius et al., 2017; Kim and Rhew, 2018; Chen et al., 2019a; Ujiie et al., 2020), AE extraction(Munkhdalai et al., 2018; Wu et al., 2021; Portelli et al., 2021; Zhang et al., 2021; Yahya and Asiri, 2022; Li et al., 2024) and AE normalisation (or mapping) (Emadzadeh et al., 2018; Combi et al., 2018). Some papers address multiple tasks simultaneously, such as classification combined with extraction (Shen and Spruit, 2021), extraction paired with relation extraction (Florez et al., 2019), or a combination of classification, extraction and RE (Yang et al., 2019). Others focus on the severity of AEs and the relationship between AEs and their seriousness (Lavertu et al., 2021; D'Oosterlinck et al., 2023).

The training of classification, extraction, and normalisation models relies on the availability of a consistently reliable annotated corpus, which remains a fundamental priority. Most of the datasets used come from publicly available resources, primarily sourced from social media platforms and shared tasks. Social media platforms, particularly Twitter, have emerged as a rich source of AE-related data due to users' frequent sharing of personal experiences with drug-related side effects (Ellenius et al., 2017; Portelli et al., 2021; Zhang et al., 2021; Guellil et al., 2022; Emadzadeh et al., 2018; Li et al., 2024). The Social Media Mining for Health (SMM4H) dataset, which contains tweets annotated for AE classification, extraction, and normalisation, and other shared task datasets (such as MADE and n2c2) has become one of the most widely used resources in AE detection studies (Yang et al., 2019; Florez et al., 2019; Chen et al., 2019b; Portelli et al., 2021). In addition to social media data, clinical datasets such as SIDER[1], FAERS[2], VAERS[3] (or KAERS for the Korean system) or those from pharmacovigilance databases (in English or other languages) (Kim and Rhew, 2018; Combi et al., 2018; Chen et al., 2019a; Ujiie et al., 2020; Shen and Spruit, 2021; Lavertu et al., 2021; Li et al., 2024) have also been widely used in AE-related research. However, we also observe a scarcity of studies that focus on electronic health records (Yang et al., 2019; Wu et al., 2021) for the extraction of AEs.

The majority of datasets used in previous studies have been manually annotated, often involving two or three annotators to ensure reliability (Tafti et al., 2017; Emadzadeh et al., 2018; Ujiie

[1] SIDER contains information on marketed medicines and their recorded adverse drug reactions

[2] FAERS: FDA Adverse Event Reporting System

[3] Vaccine Adverse Event Reorting System

et al., 2020; Shen and Spruit, 2021). Inter-annotator agreement (IAA) is commonly measured using Cohen's kappa (Cohen, 1960) for classification tasks (Tafti et al., 2017; Chen et al., 2019a; Ujiie et al., 2020), while the F1-score is predominantly used for evaluating extraction tasks (Munkhdalai et al., 2018; D'Oosterlinck et al., 2023). Only a limited number of studies consider more complex annotation challenges, such as discontinuous and overlapping entities (Kim and Rhew, 2018; Yang et al., 2019), which can significantly impact the quality and usability of annotated data. Furthermore, the consistency of manual annotation largely depends on the availability and clarity of the annotation guidelines. Despite their crucial role in ensuring coherence, only a few studies explicitly present these guidelines (Henry et al., 2020; Chen et al., 2019a; Yang et al., 2019), making it difficult to assess the reproducibility and robustness of the annotation process.

2.1 Contribution

The use of NLP for detecting AEs has advanced significantly, with notable improvements in classification, Named Entity Recognition (NER), normalization and relation Extraction (RE). Transformer-based models, such as BERT and BioBERT, have played a key role in these advancements, enhancing performance across various AE-related tasks. However, challenges persist, particularly in addressing the varying severity of AEs. In clinical practice, not all AEs carry the same level of urgency - some require immediate medical intervention, while others, though less critical, still have a considerable impact. For instance, conditions like intracranial haemorrhage demand urgent attention, whereas issues such as constipation, while important, are less pressing.

To address these challenges, our work focuses on detecting 14 specific AEs, categorized by their urgency. These range from life-threatening events, such as intracranial haemorrhage and lower respiratory tract infections, to less critical conditions. We manually annotated a corpus of 2,040 discharge summaries (provided by Dataloch[4]) to identify these AEs, ensuring both accuracy and consistency through a carefully developed annotation guideline created in collaboration with clinicians. To promote further research in this area, both the annotation guidelines and the dataset have been made publicly available (upon request from Dataloch). Beyond AE detection, our system also predicts additional attributes, such as whether the event occurred in the hospital, whether it is affirmative or negated and other relevant details. Unlike many existing studies, we pay particular attention to discontinuous and overlapping events, which are often overlooked in previous research. To validate our dataset, we experiment with and compare a variety of model embeddings, ensuring a comprehensive evaluation of their effectiveness. For a better comparison of our contribution to the proposed studies, please refer to table 1 in Appendix A.

3 Annotation guidelines

There is no consensus on an exhaustive list of AEs. In the context of this study, we reviewed the current literature and organised a Public Partipation Involvement and Engagement (PPIE) workshop for older adults to propose a patient-orientated NLP tool. Based on the information collected, we collaborated with an expert panel of clinicians using a consensus approach to select 14 AEs (detailed in section 3.1). We selected these AEs because they are both common and have a significan impact on patients' quality of life, often requiring substantial medical intervention.

3.1 Annotation labels

In the context of this work, each type of AE represents a label. The following list provides an example for each of the different types:

1. **Fall:** "Three falls in last yr"
2. **Delirium:** "On admission, she was confused. This settled to some extent with pain relief and rehydration."
3. **Pressure injury:** "Noted decubitus ulcer on the right heel"
4. **Hip fracture:** "Admitted for a left intertrochanteric fracture"
5. **Wrist fracture** "Wrist fracture shown on the X-ray."
6. **Proximal humeral fracture:** "He sustained a neck of humerus fracture on the left side a year ago"
7. **Vertebral compression fracture:** "She has

[4] Dataloch: a Scottish regional trusted research environment (TRE)

several known osteoporotic vertebral fractures"

8. **Intracranial haemorrhage:** "Right parietal subdural haematoma visible on CT"
9. **Upper gastrointestinal haemorrhage:** "Patient admitted to A&E with 3 episodes of haematemesis."
10. **Lower gastrointestinal Haemorrhage:** "She reported some PR bleed"
11. **Lower respiratory tract infection:** "New cough and wheeze during admission. Treated as hospital-acquired pneumonia"
12. **Urinary tract infection:** "CRP was raised, urine dip was positive for blood and leukocytes and Ciprofloxacin was started for UTI."
13. **Constipation:** "Patient had difficulties voiding despite laxatives. This was resolved with an enema."
14. **Seizure:** "Found by the nurse with generalised tonic-clonic episode"

3.2 Annotation attributes

Annotation attributes are additional features or characteristics associated with the tag. We use four main attributes: Negation, low confidence, in-hospital events and Diagnosis. *Negation* can be formulated explicitly with "no" or "not" but can also appear as acronyms (e.g. NAD, meaning nothing abnormal discovered) or be inferred from the text. The annotator did when possible, not include the negation word(s) in the selection of text but simply highlight the words which reference the AE and add the negation attribute to the label (eg. "She slipped but did not fall." - This should be annotated *Falls* with the negation attribute.). The *Low confidence* should be selected when the annotator is uncertain of their choice (e.g. "She was admitted with delirium, was confused in the ambulance and had visual hallucinations at home. Her daughter said she was describing children running around but they were by themselves in the house. The delirium resolved with rehydration.") - Hallucinations are rarely present in delirium but it seems to be described by the author of the letter as such. The `in-hospital_event` attribute allows to specify that the onset of AE took place in the hospital (e.g. "New cough and wheeze during admission. Treated as hospital-acquired pneumonia".

The *Diagnosis* attribute allows the annotator to further qualify the annotated text when the AE is not directly present in the patient's admission, attendance or report. For this purpose, four tags are available:

- *History of*, this can be mentioned in the patient's past medical history or something which happened before the admission or attendance (eg. "PMH: T2DM, Parkinson's disease, subdural haemorrhage." - where PMH reffers to the past medical history)
- *Suspected*, used when the writer of the clinical document suspects the patient has an AE. This attribute can also be used in annotating screening tests which require further investigations to confirm a diagnosis (eg. "This patient was admitted following a GP visit who suspects she might have a wrist fracture.")
- *Referral* qualifies when a patient is being referred to a different healthcare professional (eg."His haemoglobin remained stable and was referred to the GI team to discuss outpatient investigations of this haematemesis"
- *Implicit mention* qualifies an AE which is mentioned in the text but does not apply to the patient. This includes descriptions of the patient's family history or diagnosis given to their spouse or risks of an AE happening (e.g. risk of falls). For example: "He was given laxatives for prophylaxis of constipation" - The term prophylaxis means the patient was given laxatives as a preventative measure.

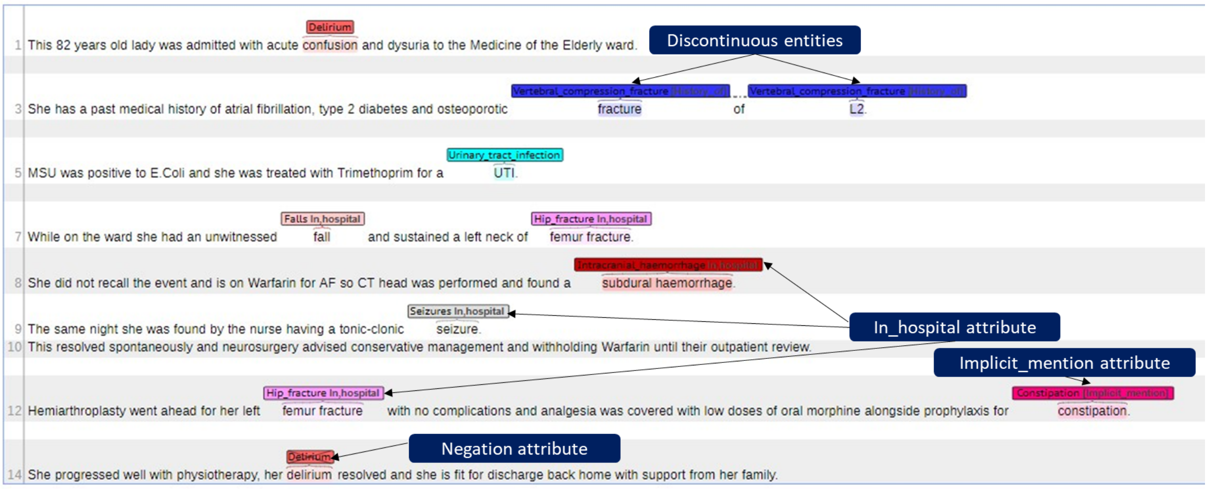
3.3 Discontinuous and overlapping entities:

Medical texts often contain distant entities which need to be annotated together to make sense of the label which is being recognised. Discontinuous entities, also named distant entities or fragmented entities, are often used for this purpose (Huang et al., 2023). For example:

- "The wrist X-ray performed in A&E confirms the presence of a fracture" - These two distant tokens should be annotated as one discontinuous entity with the label *Wrist fracture*

This does not apply when annotating two different, isolated symptoms which are associated with the same AE. For example:

- "She was confused and agitated on the ward" -



This 82 years old lady was admitted with acute confusion and dysuria to the Medicine of the Elderly ward.

She has a past medical history of atrial fibrillation, type 2 diabetes and osteoporotic fracture of L2.

MSU was positive to E.Coli and she was treated with Trimethoprim for a UTI.

While on the ward she had an unwitnessed fall and sustained a left neck of femur fracture.

She did not recall the event and is on Warfarin for AF so CT head was performed and found a subdural haemorrhage.

The same night she was found by the nurse having a tonic-clonic seizure.

This resolved spontaneously and neurosurgery advised conservative management and withholding Warfarin until their outpatient review.

Hemiarthroplasty went ahead for her left femur fracture with no complications and analgesia was covered with low doses of oral morphine alongside prophylaxis for constipation.

She progressed well with physiotherapy, her delirium resolved and she is fit for discharge back home with support from her family.

Figure 2: A synthetic example annotated in Brat includes discontinuous entities and different attributes.

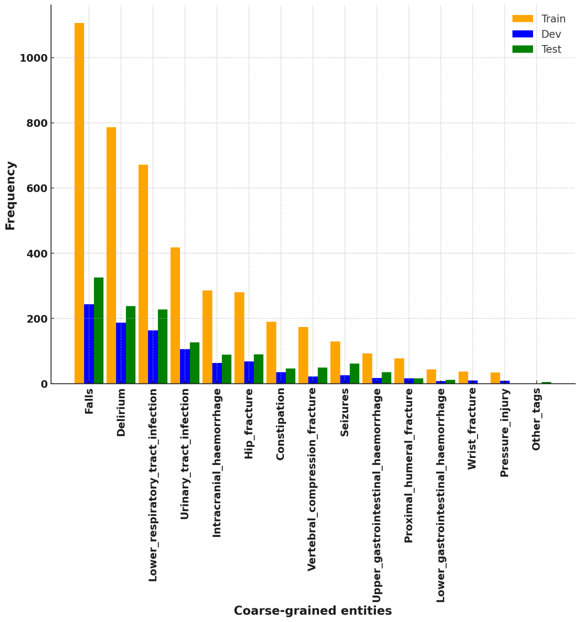


Figure 3: The distribution of coarse-grained entities in the train, dev and test datasets

since both of these symptoms, in isolation, are sufficient to detect delirium, the tokens highlighted in this example must be annotated as two separate entities with the label *Delirium*.

Medical texts also tend to contain overlapping entities which require the annotator to highlight the same element of text twice and this can also be used in combination with the distant entities (Huang et al., 2023). For example:

"When she fell, she broke her wrist and her hip" - This can be annotated with "broke" + "wrist" for the label *Wrist fracture* and "broke" + "hip" for the label *hip fracture*.

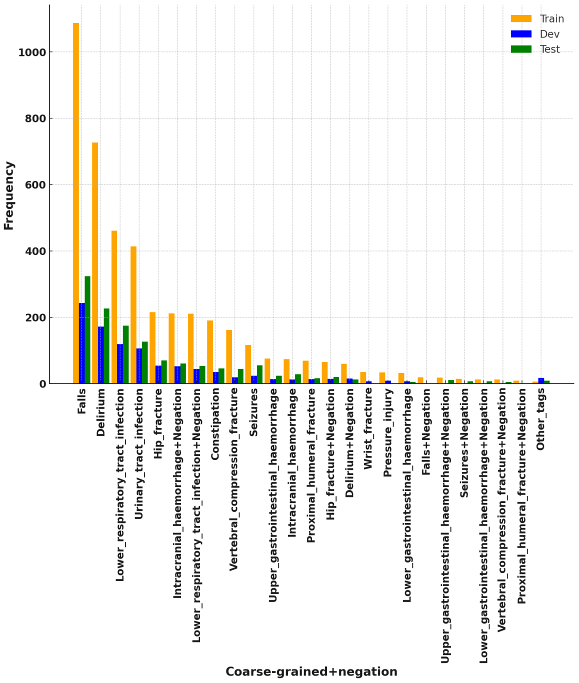


Figure 4: The distribution of coarse-grained+negation entities in the train, dev and test datasets

Due to space constraint, we present in this paper the most important points related to our guideline. The complete version of the guideline proposed to the annotators is presented in the appendix B.

4 Data annotation

We used Brat[5], a web-based text annotation tool (Stenetorp et al., 2012), to create the annotations.

[5] https://brat.nlplab.org/

Table 1: IAA between the two annotators (entity and document levels) using bratiaa and Cohen's kappa

Corpus	Annotation type	Entity level	Document level	
		F1-score	F1-score	Cohen's kappa
Pilot	Fine-grained	0.532	0.720	0.714
	coarse-grained	0.675	0.928	0.918
	coarse-grained+negation	0.675	0.911	0.905
Test	Fine-grained	0.701	0.829	0.826
	coarse-grained	0.781	0.935	0.927
	coarse-grained+negation	0.777	0.927	0.923

Figures 2, 7, 8 illustrate typical examples, showcasing some of the annotation labels and attributes defined in our annotation guidelines. Since real patient data cannot be publicly shared due to their sensitive nature, these examples are synthetics. We rely on the same examples in the section 5.3 for our error analysis part.

4.1 Dataset

We annotated data for a cohort of 200 patients, born on or before 1949, from NHS Lothian[6] hospitals. 100 individuals are aged between 70 and 79 years and 100 are aged 80 years or over. All patients had to have had more than two recent hospital visits. For this cohort, DataLoch provided us with a set of de-identified discharge summaries (2,040) belonging to these patients. Two annotators (Annotator1 and Annotator2) are involved in the annotation process. Annotator1 is a junior doctor with several years of clinical experience and Annotator2 is a geriatrician.

From the 2,040 discharge summaries, 50 documents (24,872 words) were used as a pilot set. This subset was annotated by both annotators, followed by a disagreement meeting to discuss and resolve any inconsistencies. All disagreements were settled through discussion, which also helped to clarify ambiguities and refine the annotation guidelines by adding more examples. The remaining 1,990 documents were used for training and evaluating the model. Among them, 409 documents (190,505 words) were randomly selected as the gold standard test set and annotated by the same two annotators who worked on the pilot set. The remaining 1,581 documents, used for model training (1,265 documents/612,338 words) and validation (319 documents/151,981 words), were annotated by a single annotator (Annotator1). This approach aligns with the methodology of Chen et al. (2019a), where one annotator was responsible for annotating the entire corpus, while a second annotator was involved only in annotating a subset to establish the gold standard used for the final evaluation.

4.2 Inter-annotator agreement (IAA)

In this study, we focus on both entity- and document-level annotation. At the entity level, the goal is to identify each tag along with its position (from the example shown in Figure 2, Span[9:10] *confusion → Delirium*, Span[131:133] *femur fracture → Hip_fracture (In_hospital)*). Unlike the entity-level annotation, document-level annotation does not consider entity positions or repetitions. For this example, the document-level entities include: *Delirium, Vertebral_compression_fracture (History_of), Urinary_tract_infection, etc*. We also consider three levels of annotation granularity: fine-grained, coarse-grained, and coarse-grained with negation. For fine-grained annotation, we retain all tags and attributes, resulting in labels such as *Falls+history_of+In_hospital_events*, allowing us to capture specific contextual information. In the coarse-grained annotation, we focus only on the main entities, such as *Falls*. The coarse-grained annotation with negation includes the primary entities along with a single attribute—*negation*, leading to labels like *Falls+Negation*. Figures 3 and 4 illustrate the distribution of various entities (coarse-grained and coarse-grained with negation) across the train, development, and test datasets. Due to space constraints, the fine-grained distribution is presented in the appendix C

4.2.1 Entity and document level agreements

Table 1 presents the entity- and document-level agreement scores for the pilot and test datasets, demonstrating improvements in annotation consistency. We use Bratiaa for calculating the f1-score for the entity level[7].At the entity level, fine-grained annotations showed moderate agreement (*0.532*) in the pilot corpus, while coarse-grained annotations achieved higher alignment (*0.675*). Refining annotation guidelines improved agreement in the test corpus for both fine-grained (*0.701*) and coarse-grained (*0.781*) annotations, highlighting

[6] health board in Scotland

[7] https://github.com/kldtz/bratiaa/

Table 2: Evaluation of the annotation (entity level)– comparison of the performances on 5 models

Model	Annotation type	Entity level			Document level		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Glove	Fine-grained	0.657	0.605	0.630	0.730	0.582	0.648
	Coarse-grained	0.828	0.784	0.805	0.902	0.902	0.902
	Coarse-grained+negation	0.817	0.771	0.794	0.877	0.855	0.866
BERT_uncased	Fine-grained	0.655	0.697	0.675	0.698	0.746	0.721
	Coarse-grained	0.853	0.860	0.857	**0.932**	0.946	0.939
	Coarse-grained+negation	0.600	0.576	0.588	0.715	0.666	0.689
BERT_cased	Fine-grained	0.597	0.661	0.627	0.655	0.707	0.680
	Coarse-grained	**0.865**	0.884	**0.874**	0.927	0.959	**0.943**
	Coarse-grained+negation	0.828	0.844	0.836	0.897	0.923	0.910
BioBERT	Fine-grained	0.655	0.689	0.671	0.714	0.746	0.730
	Coarse-grained	0.856	0.885	0.870	0.918	0.959	0.937
	Coarse-grained+negation	0.854	0.859	0.856	0.908	0.930	0.919
Bio-Clinical_BERT	Fine-grained	0.576	0.586	0.581	0.620	0.572	0.595
	Coarse-grained	0.840	**0.887**	0.863	0.908	**0.961**	0.933
	Coarse-grained+negation	0.816	0.868	0.841	0.874	0.947	0.909

the impact of task complexity on agreement scores. At the document level, the test corpus consistently outperformed the pilot across all annotation types, with fine-grained (*0.829* vs. *0.720*), coarse-grained (*0.935* vs. *0.928*), and coarse-grained with negation (*0.927* vs. *0.911*) annotations showing improvement. These results confirm the effectiveness of pilot studies in refining annotation guidelines and enhancing annotation consistency, particularly in document-level extraction, which is crucial for clinical applications. More detailled agreements for some entities are presented in appendix D

5 Evaluation of the annotation on the test corpus

5.1 Methodology

All models were implemented using the FlairNLP framework (Akbik et al., 2019), following a standard LSTM-CRF architecture. Each model leveraged one of five types of pre-trained embeddings: GloVe (Pennington et al., 2014), BERT (cased and uncased) (Devlin et al., 2019), BioBERT (Lee et al., 2020), or BioClinicalBERT (Alsentzer et al., 2019). These embeddings were passed into a bidirectional LSTM (Hochreiter, 1997) encoder consisting of two recurrent layers with a hidden size of 64 and a dropout rate of 0.16. A Conditional Random Field (CRF) (Lafferty et al., 2001) layer was used for the final sequence labeling task. The models were trained for 200 epochs using stochastic gradient descent (SGD) with a learning rate of 0.037 and a mini-batch size of 2, constrained by GPU availability within the Trusted Research Environment (TRE). As a result, each model took approximately 36 hours to complete training. Due to the computational limitations and runtime restrictions in the TRE, we did not perform hyperparameter optimisation. Additionally, to support the annotation of discontinuous and overlapping entities, we adapted preprocessing methods from Dai et al. (2020) to convert Brat annotations into a compatible CoNLL format [8]. Next, we generated three versions of the dataset from each CoNLL file: fine-grained (retaining all transformed entities and attributes), coarse-grained (keeping only the main entities) and coarse-grained with negation (including all entities along with the negation attribute). Each of these versions was then used to train models with five different embeddings: Glove , BERT (Devlin, 2018) uncased, BERT cased, BioBERT (Lee et al., 2020) and BioClinicalBERT (Alsentzer et al., 2019).

5.2 Entity and document level results

Table 2 presents the entity and document level evaluations of the five models used across fine-grained, coarse-grained, and coarse-grained with negation annotations. Coarse-grained annotations consistently outperformed fine-grained annotations across all models, as expected due to fewer classes and more training examples. For the entity level evaluation, BERT (cased) achieved the highest performance for coarse-grained annotations (*F1 = 0.874*), followed closely by BioBERT (*F1 =*

[8] `https://universaldependencies.org/format.html`

Table 7: Some examples

Example	Annotation type	Entity level	Document level
Example 1 (Figure 2)	Fine-grained	**Span[9:10]**:confusion/Delirium(1.0)	Delirium Vertebral_compression_fracture+Diagnosis Vertebral_compression_fracture+History Urinary_tract_infection Falls Hip_fracture Intracranial_haemorrhage+Diagnosis, Intracranial_haemorrhage+History_of Seizures Constipation
		Span[35:38]: fracture of L2/Vertebral_compression_fracture+ Diagnosis+History_of(0.8957)	
		Span[52:53]: UTI/Urinary_tract_infection(0.9993)	
		Span[62:63]: fall/Falls(0.9973)	
		Span[69:71]: femur fracture/Hip_fracture+Diagnosis+History_of(0.4575)	
		Span[92:94]: subdural haemorrhage/Intracranial_haemorrhage(0.7101)	
		Span[107:108]: seizure/Seizures(0.9212)	
		Span[131:133]: femur fracture/Hip_fracture+Diagnosis+History_of(0.8819)	
		Span[149:150]: constipation(0.9975)	
		Span[158:159]: delirium/Delirium(0.9978)	
	Coarse-grained	**Span[9:10]**:confusion/Delirium(1.0)	Delirium Vertebral_compression_fracture Urinary_tract_infection Falls Hip_fracture Intracranial_haemorrhage Seizures Constipation
		Span[35:38]: fracture of L2/Vertebral_compression_fracture(0.9998)	
		Span[52:53]: UTI/Urinary_tract_infection(1.0)	
		Span[62:63]: fall/Falls(1.0)	
		Span[69:71]: femur fracture/Hip_fracture(0.9999)	
		Span[92:94]: subdural haemorrhage/Intracranial_haemorrhage(0.9997)	
		Span[107:108]: seizure/Seizures(0.9997)	
		Span[131:133]: femur fracture/Hip_fracture(0.8958)	
		Span[149:150]: constipation/Constipation(0.9999)	
		Span[158:159]: delirium/Delirium(1.0)	
	Coarse-grained+negation	**Span[9:10]**:confusion/Delirium(1.0)	Delirium Vertebral_compression_fracture Urinary_tract_infection Falls Hip_fracture Intracranial_haemorrhage Seizures Constipation
		Span[35:38]: fracture of L2/Vertebral_compression_fracture(0.953)	
		Span[52:53]: UTI/Urinary_tract_infection(1.0)	
		Span[62:63]: fall/Falls(1.0)	
		Span[69:71]: femur fracture/Hip_fracture(0.9996)	
		Span[92:94]: subdural haemorrhage/Intracranial_haemorrhage(0.9972)	
		Span[107:108]: seizure/Seizures(0.944)	
		Span[131:133]: femur fracture/Hip_fracture(0.9986)	
		Span[149:150]: constipation/Constipation(0.9965)	
		Span[158:159]: delirium/Delirium(0.9999)	

0.870). When negation attributes were included, these models maintained strong performance, with BERT (cased) scoring *0.836* and BioBERT scoring *0.856*, demonstrating their robustness in handling annotation complexity. Bio-Clinical BERT also performed well for negation detection (*F1 = 0.841*). The results also indicate that document-level extraction benefits significantly from coarse-grained annotations, with models achieving high F1-scores. BERT (cased) performed best (*F1 = 0.943*), surpassing BioBERT (*F1 = 0.937*). These findings highlight the effectiveness of transformer-based models, particularly BERT variants, in leveraging contextual information for document-level tasks.

Fine-grained annotations posed greater challenges, yielding lower scores across all models. BioBERT and BERT (uncased) performed best for fine-grained annotations, with F1-scores of *0.730* and *0.721*, respectively (for document level). However, GloVe and Bio-Clinical BERT struggled, reflecting the increased complexity and ambiguity of fine-grained annotations. The highest F1-score for fine-grained annotations was *0.675*, achieved by BERT (uncased), followed by BioBERT (*F1 = 0.671*), for the entity-level. GloVe and Bio-Clinical BERT performed the worst, with Bio-Clinical BERT scoring below *0.630*, highlighting the impact of annotation granularity on model performance.

5.3 Error analysis

To evaluate the performance of our models, we apply them to some synthetic examples, which were created as part of our annotation guideline development. Table 4 displays the different AEs extracted (at both entity and document levels) for each level of granularity we considered for the our presented example in Figure 2 (We focus on just one example because of space constraint but four other examples are presented in Appendix E). For each granularity type, we use the best-performing model: BioBERT for entity-level and BERT_uncased for document-level in fine-grained analysis, BERT_cased for both levels in coarse-grained analysis, and BioBERT for both levels in coarse-grained+negation analysis.

The tagging of this example aligns well with the presented results, as all coarse-grained entities were successfully detected at both the entity and document levels. However, the models exhibited difficulty in identifying the In-hospital and implicit_mention attributes, as well as in detecting the negation of *delirium*. This can be primarily attributed to the limited number of instances of these categories in the training data. For instance, the dataset contains only six instances of *Seizures_in-hospital* and nine instances of *Constipation_in-hospital*, which limits the model's ability to learn these patterns effectively. Moreover, as the focus of this study is to highlight the dataset and annotation guidelines, we did not explore techniques such as data balancing or adjusting class weights to mitigate the effects of underrepresented categories. Addressing these limitations remains an important objective for future work.

6 Conclusion

This study presents a manually annotated dataset for extracting Adverse Events (AEs) from clinical text, specifically focusing on discharge summaries of elderly patients. While our work does not propose a novel method or architecture for language analysis, its primary contribution lies in the development of a high-quality, richly annotated dataset tailored for a critical real-world application: the extraction of adverse events (AEs) from clinical discharge summaries of elderly patients — a population often underrepresented in existing corpora. The novelty of our work resides in the annotation schema, which was carefully designed in collaboration with clinicians to capture clinically meaningful entities and attributes. It supports complex linguistic phenomena such as discontinuous and overlapping entities, as well as a multi-attribute tagging system (e.g., negation, suspected, referral, in-hospital), which are rarely addressed in previous AE-related datasets. Additionally, this resource enables more nuanced model development and benchmarking for clinical NLP, while also surfacing real-world challenges like entity imbalance, fine-grained annotation complexity, and the computational limitations inherent in Trusted Research Environments (TREs). Beyond fine-grained extraction, our dataset is also compatible with research tasks involving more generic annotation schemes, such as simply labeling the presence or absence of adverse events at the document level — which aligns with existing efforts like the n2c2 corpus (a subset of MIMIC annotated for AEs)

This dataset has also been extended for the detection of geriatric syndromes (Guellil et al., 2024a) and the social context of patients. Additionally, as future work, we plan to assess cross-domain generalisation by applying our annotation guidelines to the MIMIC-IV dataset (Johnson et al., 2023). Inspired by the approach in (Dai et al., 2024), which introduced a multi-domain benchmark for AE detection, this evaluation will help determine how well models trained on Dataloch generalize to MIMIC-IV, and vice versa, further advancing AE detection in clinical settings.

7 Limitations

This study was conducted within a Trusted Research Environment (TRE), which imposes certain constraints that impacted our model choices and approach including a focus on efficient model architectures, and it was not possible to deploy generative AI models (such as GPT-4 or Mistral). A key challenge in this study was addressing the imbalance of AE information in the dataset. As shown in the appendix with detailed results for each entity, performance was generally higher for more frequent entities such as falls and delirium, while less common entities showed lower performance scores. This pattern could potentially be addressed through balancing techniques or class weighting approaches, though resource constraints meant we were unable to explore these methods within the scope of this study. Similarly, hyperparameter optimisation was not feasible given the computational requirements.

Despite these challenges, working with a TRE provides invaluable access to real patient data, making it a crucial resource for clinical NLP research. While state-of-the-art techniques may require adaptations to fit within TRE infrastructure, they remain essential for ensuring the security and ethical use of sensitive health.

Acknowledgment

This research has been assisted by the DataLoch service (reference: DL_2021_009) and received favourable ethical opinion under DataLoch's delegated Research Ethics authority (Reference: 17/NS/0027). DataLoch enables access to deidentified extracts of health care data from the South-East Scotland region to approved applicants: dataloch.org.

For MIMIC IV, the collection of patient information and creation of the research resource was reviewed by the Institutional Review Board at the Beth Israel Deaconess Medical Center, who granted a waiver of informed consent and approved the data sharing initiative.

This study/project (AIM-CISC) was funded by the National Institute for Health Research (NIHR). Artificial Intelligence and Multimorbidity: Clustering in Individuals, Space and Clinical Context (AIM-CISC) grant NIHR202639. 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 research was also funded by the Legal & General Group (research grant to establish the independent Advanced Care Research Centre at the University of Edinburgh). The funder had no role in 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 Legal & General.

As this work is being extended to MIMIC IV, the authors would like to express their sincere gratitude to all the annotators involved in this annotation, including Scott Prentice, Mike Holder, Rose Penfold and Lauren Ng. Finally, the authors would like to warmly thank Richard Tobin (School of Informatics) for his amazing help in setting up Brat. They would also like to thank the ACRC management team for their valuable support.

References

B Abrahamsen, T Van Staa, R Ariely, M Olson, and C Cooper. 2009. Excess mortality following hip fracture: a systematic epidemiological review. *Osteoporosis international*, 20:1633–1650.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Daniela Alexandru and William So. 2012. Evaluation and management of vertebral compression fractures. *The Permanente Journal*, 16(4):46.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

G Ross Baker, Peter G Norton, Virginia Flintoft, Régis Blais, Adalsteinn Brown, Jafna Cox, Ed Etchells, William A Ghali, Philip Hébert, Sumit R Majumdar, et al. 2004. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. *Cmaj*, 170(11):1678–1686.

J Alfredo Caceres and Joshua N Goldstein. 2012. Intracranial hemorrhage. *Emergency medicine clinics of North America*, 30(3):771–794.

Michael Camilleri, Alexander C Ford, Gary M Mawe, Phil G Dinning, Satish S Rao, William D Chey, Magnus Simrén, Anthony Lembo, Tonia M Young-Fadok, and Lin Chang. 2017. Chronic constipation. *Nature reviews Disease primers*, 3(1):1–19.

Jinying Chen, John Lalor, Weisong Liu, Emily Druhl, Edgard Granillo, Varsha G Vimalananda, and Hong Yu. 2019a. Detecting hypoglycemia incidents reported in patients' secure messages: using cost-sensitive learning and oversampling to reduce data imbalance. *Journal of medical Internet research*, 21(3):e11990.

Tao Chen, Mark Dredze, Jonathan P Weiner, Leilani Hernandez, Joe Kimura, Hadi Kharrazi, et al. 2019b. Extraction of geriatric syndromes from electronic health record clinical notes: assessment of statistical natural language processing methods. *JMIR medical informatics*, 7(1):e13039.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Carlo Combi, Margherita Zorzi, Gabriele Pozzani, Elena Arzenton, and Ugo Moretti. 2018. Normalizing spontaneous reports into meddra: Some experiments with magicoder. *IEEE Journal of Biomedical and Health Informatics*, 23(1):95–102.

Steven R Cummings, Jennifer L Kelsey, Michael C Nevitt, and KENNETH J O'DOWD. 1985. Epidemiology of osteoporosis and osteoporotic fractures. *Epidemiologic reviews*, 7(1):178–208.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. An effective transition-based model for discontinuous ner. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5860–5870.

Xiang Dai, Sarvnaz Karimi, Abeed Sarker, Ben Hachey, and Cecile Paris. 2024. Multiade: A multi-domain benchmark for adverse drug event extraction. *Journal of Biomedical Informatics*, 160:104744.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186.

Ian P Donald, Roger G Smith, John G Cruikshank, Robert A Elton, and Margaret E Stoddart. 1985. A study of constipation in the elderly living at home. *Gerontology*, 31(2):112–118.

Karel D'Oosterlinck, François Remy, Johannes Deleu, Thomas Demeester, Chris Develder, Klim Zaporojets, Aneiss Ghodsi, Simon Ellershaw, Jack Collins, and Christopher Potts. 2023. Biodex: Large-scale biomedical adverse drug event extraction for real-world pharmacovigilance. *arXiv preprint arXiv:2305.13395*.

Laura E Edsberg, Joyce M Black, Margaret Goldberg, Laurie McNichol, Lynn Moore, and Mary Sieggreen. 2016. Revised national pressure ulcer advisory panel pressure injury staging system: revised pressure injury staging system. *Journal of Wound, Ostomy, and Continence Nursing*, 43(6):585.

Johan Ellenius, Lucie M Gattepaille, Sara Vidlin, Carrie Pierce, and Tomas Bergvall. 2017. Detecting and encoding mentions of suspected adverse events in twitter using natural language processing. In *PHARMACOEPIDEMIOLOGY AND DRUG SAFETY*, volume 26, pages 36–37. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA.

Ehsan Emadzadeh, Abeed Sarker, Azadeh Nikfarjam, and Graciela Gonzalez. 2018. Hybrid semantic analysis for mapping adverse drug reaction mentions in tweets to medical terminology. In *AMIA Annual Symposium Proceedings*, volume 2017, page 679.

JJ Farrell and LS Friedman. 2005. the management of lower gastrointestinal bleeding. *Alimentary pharmacology & therapeutics*, 21(11):1281–1298.

Edson Florez, Frederic Precioso, Romaric Pighetti, and Michel Riveill. 2019. Deep learning for identification of adverse drug reaction relations. In *Proceedings of the 2019 International Symposium on Signal Processing Systems*, pages 149–153.

Kevin A Ghassemi and Dennis M Jensen. 2013. Lower gi bleeding: epidemiology and management. *Current gastroenterology reports*, 15:1–6.

Assaf Gottlieb, Robert Hoehndorf, Michel Dumontier, and Russ B Altman. 2015. Ranking adverse drug reactions with crowdsourcing. *Journal of medical Internet research*, 17(3):e80.

James F Griffith. 2015. Identifying osteoporotic vertebral fracture. *Quantitative imaging in medicine and surgery*, 5(4):592.

Imane Guellil, Salomé Andres, Bruce Guthrie, Atul Anand, Huayu Zhang, Abul Kalam Hasan, Honghan Wu, and Beatrice Alex. 2024a. Enhancing natural language processing capabilities in geriatric patient care: An annotation scheme and guidelines. In *International Conference on Applications of Natural Language to Information Systems*, pages 207–217. Springer.

Imane Guellil, Jinge Wu, Aryo Pradipta Gema, Farah Francis, Yousra Berrachedi, Nidhaleddine Chenni, Richard Tobin, Clare Llewellyn, Stella Arakelyan, Honghan Wu, Bruce Guthrie, and Beatrice Alex. 2024b. Natural language processing for detecting adverse drug events: A systematic review protocol. *NIHR Open Research*, 3:67.

Imane Guellil, Jinge Wu, Honghan Wu, Tony Sun, and Beatrice Alex. 2022. Edinburgh_ucl_health@ smm4h'22: From glove to flair for handling imbalanced healthcare corpora related to adverse drug events, change in medication and self-reporting vaccination. In *Proceedings of COLING. International conference on computational Linguistics*, volume 2022, page 148.

Andrew J Hall, Nicholas D Clement, Hani Abdul-Jabar, Rashid Abu-Rajab, Ahmed Abugarja, Karen Adam, Héctor J Aguado Hernández, Gedeón Améstica Lazcano, Sarah Anderson, Mahmood Ansar, et al. 2022. Impact-global hip fracture audit: Nosocomial infection, risk prediction and prognostication, minimum reporting standards and global collaborative audit: Lessons from an international multicentre study of 7,090 patients conducted in 14 nations during the covid-19 pandemic. *The Surgeon*, 20(6):e429–e446.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

S Hochreiter. 1997. Long short-term memory. *Neural Computation MIT-Press*.

Peixin Huang, Xiang Zhao, Minghao Hu, Zhen Tan, and Weidong Xiao. 2023. T 2-ner: At wo-stage span-based framework for unified named entity recognition with t emplates. *Transactions of the Association for Computational Linguistics*, 11:1265–1282.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

Sarabjot Kaur, Ravinder Garg, Simmi Aggarwal, Sumit Pal Singh Chawla, and Ranabir Pal. 2018. Adult onset seizures: Clinical, etiological, and radiological profile. *Journal of family medicine and primary care*, 7(1):191.

Hyon Hee Kim and Ki Yon Rhew. 2018. A machine learning approach to classification of case reports on adverse drug reactions using text mining of expert opinions. In *Advances in Computer Science and Ubiquitous Computing: CSA-CUTE 17*, pages 1072–1077. Springer.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Adam Lavertu, Tymor Hamamsy, and Russ B Altman. 2021. Quantifying the severity of adverse drug reactions using social media: network analysis. *Journal of Medical Internet Research*, 23(10):e27714.

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.

Sun H Lee, Patricia Dargent-Molina, and Gérard Bréart. 2002. Risk factors for fractures of the proximal humerus: results from the epidos prospective study. *Journal of Bone and Mineral research*, 17(5):817–825.

Yiming Li, Deepthi Viswaroopan, William He, Jianfu Li, Xu Zuo, Hua Xu, and Cui Tao. 2024. Improving entity recognition using ensembles of deep learning and fine-tuned large language models: a case study on adverse event extraction from multiple sources. *arXiv preprint arXiv:2406.18049*.

Evelyn Lo, Lindsay E Nicolle, Susan E Coffin, Carolyn Gould, Lisa L Maragakis, Jennifer Meddings, David A Pegues, Ann Marie Pettis, Sanjay Saint, and Deborah S Yokoe. 2014. Strategies to prevent catheter-associated urinary tract infections in acute care hospitals: 2014 update. *Infection Control & Hospital Epidemiology*, 35(5):464–479.

Alasdair MJ MacLullich, Anna Beaglehole, Roanna J Hall, and David J Meagher. 2009. Delirium and long-term cognitive impairment. *International review of psychiatry*, 21(1):30–42.

Jay Magaziner, Eleanor M Simonsick, T Michael Kashner, J Richard Hebel, and John E Kenzora. 1990. Predictors of functional recovery one year following hospital discharge for hip fracture: a prospective study. *Journal of gerontology*, 45(3):M101–M107.

Darshini Mahendran and Bridget T McInnes. 2021. Extracting adverse drug events from clinical notes. *AMIA Summits on Translational Science Proceedings*, 2021:420.

Melissa LP Mattison. 2020. Delirium. *Annals of internal medicine*, 173(7):ITC49–ITC64.

Tsendsuren Munkhdalai, Feifan Liu, Hong Yu, et al. 2018. Clinical relation extraction toward drug safety surveillance using electronic health record narratives: classical learning versus deep learning. *JMIR public health and surveillance*, 4(2):e9361.

Rachel M Murphy, Joanna E Klopotowska, Nicolette F de Keizer, Kitty J Jager, Jan Hendrik Leopold, Dave A Dongelmans, Ameen Abu-Hanna, and Martijn C Schut. 2023. Adverse drug event detection using natural language processing: A scoping review of supervised learning methods. *Plos one*, 18(1):e0279842.

Andrew M Naidech. 2011. Intracranial hemorrhage. *American journal of respiratory and critical care medicine*, 184(9):998–1006.

Michael C Nevitt, Steven R Cummings, and Study of Osteoporotic Fractures Research Group. 1993. Type of fall and risk of hip and wrist fractures: the study of osteoporotic fractures. *Journal of the American Geriatrics Society*, 41(11):1226–1234.

Christine Norton. 2006. Constipation in older patients: effects on quality of life. *British Journal of Nursing*, 15(4):188–192.

Elizabeth A O'Keefe, Nicholas J Talley, Alan R Zinsmeister, and Steven J Jacobsen. 1995. Bowel disorders impair functional status and quality of life in the elderly: a population-based study. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 50(4):M184–M189.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Beatrice Portelli, Daniele Passabì, Edoardo Lenzi, Giuseppe Serra, Enrico Santus, and Emmanuele Chersoni. 2021. Improving adverse drug event extraction with spanbert on different text typologies. In *International Workshop on Health Intelligence*, pages 87–99. Springer.

Adnan I Qureshi, Stanley Tuhrim, Joseph P Broderick, H Hunt Batjer, Hideki Hondo, and Daniel F Hanley. 2001. Spontaneous intracerebral hemorrhage. *New England Journal of Medicine*, 344(19):1450–1460.

Samarth Rawal, Siddharth Rawal, Saadat Anwar, and Chitta Baral. 2019. Identification of adverse drug reaction mentions in tweets–smm4h shared task 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 136–137.

LR Schiller. 2001. The therapy of constipation. *Alimentary pharmacology & therapeutics*, 15(6):749–763.

M-A Schürch, René Rizzoli, Bernadette Mermillod, Harold Vasey, Jean-Pierre Michel, and J-P Bonjour. 1996. A prospective study on socioeconomic aspects of fracture of the proximal femur. *Journal of Bone and Mineral Research*, 11(12):1935–1942.

Zhengru Shen and Marco Spruit. 2021. Automatic extraction of adverse drug reactions from summary of product characteristics. *Applied Sciences*, 11(6):2663.

American Thoracic Society, Infectious Diseases Society of America, et al. 2005. Guidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia. *American journal of respiratory and critical care medicine*, 171(4):388.

Adrian J Stanley and Loren Laine. 2019. Management of acute upper gastrointestinal bleeding. *Bmj*, 364.

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.

Daniel A Sterling, Judith A O'connor, and John Bonadies. 2001. Geriatric falls: injury severity is high and disproportionate to mechanism. *Journal of Trauma and Acute Care Surgery*, 50(1):116–119.

Ofelia C Tablan, Larry J Anderson, Richard E Besser, Carolyn B Bridges, and Rana A Hajjeh. 2003. Guidelines for preventing health-care-associated pneumonia, 2003; recommendations of cdc and the healthcare infection control practices advisory committee.

Ahmad P Tafti, Jonathan Badger, Eric LaRose, Ehsan Shirzadi, Andrea Mahnke, John Mayer, Zhan Ye, David Page, Peggy Peissig, et al. 2017. Adverse drug event discovery using biomedical literature: a big data neural network adventure. *JMIR medical informatics*, 5(4):e9170.

John M Tallon, Stacy Ackroyd-Stolarz, Saleema A Karim, and David B Clarke. 2008. The epidemiology of surgically treated acute subdural and epidural hematomas in patients with head injuries: a population-based study. *Canadian journal of surgery*, 51(5):339.

Shogo Ujiie, Shuntaro Yada, Shoko Wakamiya, Eiji Aramaki, et al. 2020. Identification of adverse drug event–related japanese articles: natural language processing analysis. *JMIR Medical Informatics*, 8(11):e22661.

ME Van Leerdam. 2008. Epidemiology of acute upper gastrointestinal bleeding. *Best practice & research Clinical gastroenterology*, 22(2):209–224.

Itziar Vergara, Kalliopi Vrotsou, Miren Orive, Susana Garcia-Gutierrez, Nerea Gonzalez, Carlota Las Hayas, and Jose M Quintana. 2016. Wrist fractures and their impact in daily living functionality on elderly people: a prospective cohort study. *BMC geriatrics*, 16:1–8.

Charles Vincent, Graham Neale, and Maria Woloshynowych. 2001. Adverse events in british hospitals: preliminary retrospective record review. *Bmj*, 322(7285):517–519.

Florian ME Wagenlehner, Mete Cek, Kurt G Naber, Hiroshi Kiyota, and Truls E Bjerklund-Johansen. 2012. Epidemiology, treatment and prevention of healthcare-associated urinary tract infections. *World journal of urology*, 30:59–67.

William E Whitehead, Donald Drinkwater, Lawrence J Cheskin, Barbara R Heller, and Marvin M Schuster. 1989. Constipation in the elderly living at home: definition, prevalence, and relationship to lifestyle and health status. *Journal of the American Geriatrics Society*, 37(5):423–429.

Hong Wu, Jiatong Ji, Haimei Tian, Yao Chen, Weihong Ge, Haixia Zhang, Feng Yu, Jianjun Zou, Mitsuhiro Nakamura, Jun Liao, et al. 2021. Chinese-named entity recognition from adverse drug event records: radical embedding-combined dynamic embedding–based bert in a bidirectional long short-term conditional random field (bi-lstm-crf) model. *JMIR medical informatics*, 9(12):e26407.

Anwar Ali Yahya and Yousef Asiri. 2022. Automatic detection of adverse drug reactions from online health forums. In *2022 13th International Conference on Information and Communication Systems (ICICS)*, pages 416–421. IEEE.

Xi Yang, Jiang Bian, Yan Gong, William R Hogan, and Yonghui Wu. 2019. Madex: a system for detecting medications, adverse drug events, and their relations from clinical notes. *Drug safety*, 42:123–133.

Tongxuan Zhang, Hongfei Lin, Yuqi Ren, Zhihao Yang, Jian Wang, Xiaodong Duan, and Bo Xu. 2021. Identifying adverse drug reaction entities from social media with adversarial transfer learning model. *Neurocomputing*, 453:254–262.

Appendices

A Table comparing the proposed studies to the proposed work

Table 1 provides a comprehensive summary of prior studies included in the related work section. For each study, we specify its classification, the methodological approach employed, the datasets utilized, the number of annotators involved in the annotation process, the models and machine learning algorithms applied, and whether the authors provide annotation guidelines for manually annotated corpora. Additionally, we examine whether the studies consider discontinuous and overlapping entities. Furthermore, we present the same elements for our proposed work to underscore our contribution and facilitate a clear comparison between existing research and the novel aspects of our study.

B Detailed annotation guideline provided to the annotators

B.1 Introduction

The purpose of this work is to extract Adverse Events (AE) in elderly patients' electronic health records (discharge summaries) to be used as a novel source of data for health and social care research. In order to automate this process, we rely on Natural Language Processing (NLP), a sub-field of artificial intelligence which makes use of machine learning to analyse human language. This requires the creation of a manually annotated dataset of clinical free-text on which the NLP model can be trained on. In this annotated corpus, the elements of interest (AE) are manually highlighted and given a label. The aim of this document is to provide a guideline for this manual annotation.

Note that all examples used in this document are created by clinicians in our team for the purpose of this document, they are not extracted from clinical data.

B.2 Annotation categories

Adverse Events (AE) can be defined as harmful events or undesired harmful effects caused by healthcare management rather than the patient's underlying disease. They have a wide range of severity and can result in prolonged hospital stay,

Table 1: Related work summary

Work	Category	Approach	Datasets / lexicons	Sources	Annotation guideline	Number annotators	Entities	Discontinuous entities	Overlapping	IAA used	Model embedding/ontology	ML algorithm
(Tafti et al., 2017)	Classification	Binary classification (sentence level)	Biomedical articles + social media data	Pubmed + Med-Help + WenMD + patientinfo	No	3	AEs/No-AEs	No	No	Kappa	Word2vec	BigNN + SVM, decision tree+ NB
(Ellenius et al., 2017)	Classification	Binary classification	Medical product tweets / VigiBase medical event dictionary	Twitter	No	-	suspected AEs	No	No	-	MedDRA	LR
(Chen et al., 2019a)	Classification	Binary classification	secure messages threads (patients with diabetes and healthcare provider)	U.S. Department of Veterans Affairs system	yes (simple one)	2 (one the whole dataset and the second for a subset)	Hypoglycemia / No-Hypoglycemia	No	No	Kappa	Word2vec	LR / RF / SVM with class wighting (SMOTE)
(Ujiie et al., 2020)	Classification	Binary classification	Japanese medical articles	Japanese pharmaceutical company	No	2	AE, NoAE	No	No	Kappa	bag of words, standard disease, context word tokens, etc	logistic regression
(Shen and Spruit, 2021)	Classification + extraction	IBM Watson Natural Language Understanding API	Electronic Medicines Compendium (EMC)	European Medicines Agency	No	2	ADE termes	No	No	-	-	Rule-based + IBM Watson
(Yang et al., 2019)	Classification +extraction + relation extraction	Binary classification +NER	MADE dataset	EHR from the University of Massachusetts Medical School	Yes	-	AE +drug +dose +frequency + route +duration	Yes	Yes	-	word + character level embedding	SVM + LSTM-CRF
(Kim and Rhew, 2018)	Classification	multiple class classification	Korea Adverse Event Reporting System (KAERS) database	Korea Institute of Drug Safety and Risk Management (KIDS)	No	-	Certain, Probable, Possible, Unlikely, Unclassified, Unclassifiable	No	Yes	-	TF-IDF	NB +Laplace smoothing
(Portelli et al., 2021)	Extraction	NER	SMM4H +CADEC	Twitter + AskaPatient	No	-	AEs	No	No	-	BERT +SpanBERT + BioBERT	CRF
(Wu et al., 2021)	Extraction	NER	AE reports	Jiangsu Province AE Monitoring Center +Drum Tower Hospital	No	-	Drug, reason, AE	No	No	-	BERT	Bi-LSTM +CRF
(Emadzadeh et al., 2018)	Normalisation	Entity linking	corpus of tweets +UMLS	Twitter +the U.S. National Library of Medicine (NLM)	No	3	AE, Indication(symptoms), drugs	No	No	agreement based on a third annotator	UMLS	Rule-based matching +LSA +HSA +regression model
(Combi et al., 2018)	Normalisation	Mapping	spontaneous reports	pharmacovigilance databases	No	-	AEs	No	No	-	MEDDRA	String similarity algorithms +contextual analysis
(Munkhdalai et al., 2018)	Extraction	NER +relation extraction	EHR narratives	clinical setting	No	-	AE, drugs, indication, severity	No	No	F1 score (document level)	one-hot encodding +n-grams	SVM + RNN + LSTM + Supervised rule induction
(Florez et al., 2019)	Extraction +relation extraction	NER	n2c2	MIMICIII	No	-	Drugs, AEs, Attributes (dosage, route, duration)	No	No	-	DeBERTa, MeDeBERTa, RoBERTa	Rule-based, NN
(Lavertu et al., 2021)	ADE severity	semi-supervised method	FAERS +Gottlieb severity (Gottlieb et al., 2015)	openFDA website	No	-	severity of AEs	No	No	-	RedMed +MedVRA	Lexical network +label propagation algorithms
(Zhang et al., 2021)	Extraction	NER	Twimed-pubmed	Twitter +Pubmed	No	-	AE, drugs	No	No	-	Word2vec	CharCNN
(Karimi et al., 2015)	Resource construction +normalisation	Manual annotation	CADEC	AskaPatient	Yes	4	Drug, ADR, Disease, Symptom, Finding	Yes	Yes	customised metric	SNOMED CT, MedDRA	-
(Yahya and Asiri, 2022)	Extraction	lexicon-based approach	Patient review +Consumer Health Vocabulary + SIDER	Askapatient + WebMD	No	-	AE for anitiepileptic drugs	No	No	-	No	Lexicon-based approach
(D'Oosterlinck et al., 2023)	Extraction + severity of events	dataset construction + few shot learning	BioDEX	PubMed articles + Drug Safety Reports	No	-	AE, drugs, seriousness, patient gender	No	No	F1 score	text-davinci-002, gpt-3.5, gpt-4, FLAN-T5-Large, FLAN-T5-XL	-
(Li et al., 2024)	Extraction	NER	reports + posts	VAERS + Twitter + Reddit	No	-	Vaccine, Shot, AE	No	No	-	BioBERT, GPT-2, GPT-3.5, GPT-4, Llma-2 7b, Llma-2 13b	RNN
Our work	Extraction	NER (Token + document level)	discharge summaries	Dataloch	Yes	2 (one the whole data and the other for a subset)	AEs (falls, delirium, hip-fracture, intracranial hemorrhage, lower tract infection, etc.	Yes	Yes	F1-score (token level + document level) and Kappa (document level)	Glove, BERT uncased, BERT cased, BioBERT, BIO-Clinical BERT	CRF + LSTM (FlairNLP)

permanent disability or contribute to death (Rawal et al., 2019; Baker et al., 2004; Vincent et al., 2001). Different AE are referenced in the literature including major osteoporotic fractures, intracranial haemorrhage and constipation. In this work, we focus on 14 AEs: falls, delirium, pressure injury, hip fractures, wrist fractures, proximal humeral fractures, vertebral compression fractures, intracranial haemorrhage, upper gastrointestinal haemorrhage, lower gastrointestinal haemorrhage, lower respiratory tract infection, urinary tract infection, constipation and seizures.

We present in the following part, the list of AEs we wish to extract with a short definition for each of them:

1. Falls: *are an events which result in a person coming to rest inadvertently on the ground or floor or other lower level*[9]. *Falls among the elderly, including same-level falls, are a common source of both high injury severity and mortality, much more so than in younger patients* (Sterling et al., 2001).

2. Delirium: *is an acute state of brain failure marked by sudden onset of confusion, a fluctuating course, inattention, and often an abnor-*

[9] https://www.who.int/news-room/fact-sheets/detail/falls

mal level of consciousness (Mattison, 2020). *It is generally reversible and can occur in response to a wide range of precipitants including acute illness, surgery, trauma or drugs, though a clear precipitant is not identifiable in around 10% of cases* (MacLullich et al., 2009).

3. Pressure injury: *is damage to the skin and the deeper layers of tissue under the skin caused by prolonged pressure to an area of the body. It is more likely to happen if a person has to stay in a bed or chair for a long time.*[10] *This term includes any soft tissue injury, even without ulceration* (Edsberg et al., 2016) *and it is also known as "decubitus ulcer", "bedsore" or "pressure sore".*

4. Hip fractures: *Proximal femoral fractures (including femoral neck, intertrochanteric and subtrochanteric fractures) are the most common reason for admission to an acute orthopaedic ward and are associated with excess mortality up to 36% in the first year* (Abrahamsen et al., 2009). *Over 35% of patients will not return to their pre-fracture functional status after 12 months* (Magaziner et al., 1990) *and 18% of patients require a more care-intensive environment within the following year*(Schürch et al., 1996). *This definition excludes periprosthetic fractures, isolated fractures of the greater trochanter, the pubic rami or the acetabulum* (Hall et al., 2022).

5. Wrist fractures: *These consist of a fracture of the distal end of the radius, the distal end of the ulna or both, and over 90% of them involve a fall* (Nevitt et al., 1993). *They are the most common amongst women until the age of 75, after which their frequency is surpassed by hip fractures* (Cummings et al., 1985). *Furthermore, while wrist fractures are rarely fatal and cause much less disability than hip fractures, a functional decline can be observed at 6 months post-fracture in 33% of patients over the age of 65* (Vergara et al., 2016).

6. Proximal humeral fractures: *They are the third most frequent fracture in the elderly following hip and wrist. While the long-term functional outcome is satisfactory in 80% of patients after a simple humeral fracture without displacement, displaced proximal humeral fractures may require lengthy hospitalisation and generally lead to long-term functional deficit* (Lee et al., 2002). *They include humeral fractures of the anatomical neck, the great tuberosity, the proximal end, the surgical neck and the upper epiphysis.*

7. Vertebral compression fractures: *Usually diagnosed on radiography, a vertebral fracture is confirmed when there is an approximate 20% loss in vertebral body height relative to normal looking adjacent vertebra* (Griffith, 2015). *Approximately 25% of all postmenopausal women in the US get a compression fracture during their lifetime. Osteoporosis is the most common aetiology for these fractures and they can occur during trivial events such as lifting a light object, a vigorous cough or sneeze or turning in bed. Although vertebral compression fractures rarely require hospital admission, they have the potential to cause significant disability and morbidity, often leading to incapacitating back pain for many months* (Alexandru and So, 2012).

8. Intracranial haemorrhage: *This term refers to any bleeding within the intracranial vault, including the brain parenchyma and surrounding meningeal spaces. This definition incorporates four main diagnoses based on anatomical location: epidural haematomas, subdural haematomas, subarachnoid haemorrhages and intracerebral haemorrhages (including parenchymal contusions). Intracranial haemorrhage may be spontaneous, precipitated by an underlying vascular malformation, induced by trauma, or related to therapeutic anticoagulation. Subdural and epudural haematomas are usually traumatic injuries* (Naidech, 2011). *Subdural haematomas are much more common, they occur in about 30% of severe head injuries and the underlying brain damage is usually much more severe than with epidural heamatomas* (Tallon et al., 2008). *Non-traumatic subarachnoid haemorrhages are most commonly due to the rupture of an intracranial aneurysm and for intracerebral haemorrhages, hypertension is the most important risk factor* (Qureshi et al., 2001). *Most patients who survive intracranial haemorrhage have a markedly lower quality of*

[10] https://www.nice.org.uk/guidance/cg179/ifp/chapter/what-is-a-pressure-ulcer

life than the general population. The 30-day mortality rate of spontaneous intracerebral haemorrhages ranges from 35% to 52% with only 20% of survivors expected to have full functional recovery at 6 months (Caceres and Goldstein, 2012).

9. Upper gastrointestinal haemorrhage: *It is a common medical emergency worldwide and refers to bleeding from the oesophagus, stomach, or duodenum. Patients present with haematemesis or melaena, although haematochezia can occur in the context of a major bleed and is typically associated with haemodynamic instability* (Stanley and Laine, 2019). *Peptic ulcer bleeding is the most common cause of upper gastrointestinal bleeding, responsible for about 50% of all cases, followed by oesophagitis and erosive disease. Variceal bleeding is less frequent in the general population, however, it has been found to be the cause of bleeding in cirrhotic patients in 50–60%* (Van Leerdam, 2008).

10. Lower gastrointestinal haemorrhage: *Haematochezia is the typical symptom of lower GI bleed, although as mentioned previously, some severe upper GI haemorrhages can also present with red blood in the stools. Similarly, while melaena typically indicates bleeding from a foregut location, it may result from bleeding from the small intestine or the right colon, particularly with slow GI bleeding or slow GI transit* (Ghassemi and Jensen, 2013). *The most common causes of lower gastrointestinal bleeding are diverticulosis, angiodysplasia, haemorrhoids, and ischaemic colitis* (Farrell and Friedman, 2005).

11. Lower respiratory tract infection (LRTI): *These include any bacterial, viral or fungal infection of the respiratory tree and will usually manifest as a pneumonia, bronchitis or lung abscess. LRTI developed in hospital (healthcare-associated pneumonias or hospital acquired pneumonias) are usually caused by bacteria. In these cases, attributable mortality rates of 20 to 33% have been reported* (Tablan et al., 2003; Society et al., 2005) *and its presence increases hospital stay by an average of 7 to 9 days.*

12. Urinary tract infection (UTI): *These include infections of the bladder (cystitis), urethra (urethritis) or kidneys (pyelonephritis) and are most frequently caused by bacteria. When acquired in hospital, they are the most frequent healthcare-associated infections and account for more than 40% of all healthcare-associated infections in a general hospital* (Wagenlehner et al., 2012) . *The vast majority of these are associated with the presence of an indwelling urinary catheter* (Lo et al., 2014).

13. Constipation: *Constipation is used to describe a variety of symptoms, including hard stools, excessive straining or infrequent bowel movements. Many disorders can cause constipation including neurologic disease (e.g. spinal cord injury, Parkinson's disease or multiple sclerosis), rectoanal problems (e.g. anal strictures, proctitis), iatrogenic conditions (e.g. drugs or previous surgeries), endocrine and metabolic diseases (e.g. diabetes or hypothyroidism), lifestyle factors such as lack of mobility and dietary factors, namely low residue diets* (Camilleri et al., 2017; Schiller, 2001). *A number of studies have shown that constipation is significantly associated with reductions in elderly patients' quality of life and functional status* (Norton, 2006; O'Keefe et al., 1995; Donald et al., 1985; Whitehead et al., 1989).

14. Seizures: *Seizures involve sudden, temporary, bursts of electrical activity in the brain that change or disrupt the way messages are sent between brain cells. These electrical bursts can cause involuntary changes in body movement or function, sensation, behaviour or awareness*[11]. *They can happen in the context of epilepsy or, be provoked by an acute medical illness. In the case of provoked seizures, the main causes are trauma, central nervous system infections, space-occupying lesions, cerebrovascular accidents, metabolic disorders, and drugs* (Kaur et al., 2018).

B.3 General instructions

B.3.1 Background

Named entity recognition is the NLP method we are employing to extract structured information from our unstructured data. This task aims to precisely locate and classify the AE within the clinical free text documents by assigning tokens with a

[11] `https://www.epilepsy.com/what-is-epilepsy/understanding-seizures`

given label.

Tokens are units of text, usually words or abbreviations and are separated by punctuation marks such as spaces, commas, full stops, slashs, hyphen, brackets or numbers.

Labels are used in many different domains, the most common example would be their use for de-identification: automatically detecting and redacting the names of persons, organisations or locations from sensitive documents. For example:

- Mark Elliot Zuckerberg is the founder of Facebook. He lives in LA.

In this example, the label *Person* should be associated with the token "Mark Elliot Zuckerberg", the label *Organisation* should be associated with the token "Facebook" and the label *Location* should be associated with the token "LA". This demonstrates that an entity could be extracted from a single word (i.e. Facebook), a set of words (i.e. Mark Elliot Zuckerberg) or an acronym (i.e. LA).

Annotators must be careful, however, when selecting multiple tokens, to always aim for as few words as possible while retaining the meaning of the label they are tagging. This rigorous process is essential to the training of the language model and will also allow a more uniform annotation between different clinicians.

The de-identification example given above is rather straightforward with a single word or a short string of adjacent words matching one label. Medical text is often more complex: terms can be fragmented and non-adjacent and some entities can overlap, meaning that one word can describe more than one label. Discontinuous entities and overlapping entities are detailed further below.

B.3.2 Discontinuous entities:

Medical texts often contain distant entities which need to be annotated together to make sense of the label which is being recognised. Discontinuous entities, also named distant entities or fragmented entities, are often used for this purpose (Huang et al., 2023). For example:

- “The wrist X-ray performed in A&E confirms the presence of a fracture” - These two distant tokens should be annotated as one discontinuous entity with the label *Wrist fracture*

When deciding to extract discontinuous entities, the annotator should aim to select terms which are as close as possible to each other in the text while retaining the meaning of the label, as shown in the examples below:

- "The X-ray of her left femur showed degenerative changes of her hip joint and a displaced extracapsular fracture" - The highlighted text is given the label *proximal femoral fracture* (or *hip fracture*). In this example, rather than using the term "femur", the annotator should aim to reduce the distance between the two elements of the text which form the fragmented entity and select "hip" which is closer to the word "fracture".
- “The wrist X-ray performed in A&E confirms the presence of a fracture of the radius” - The highlighted text is given the label *wrist fracture* and instead of selecting the word "wrist", the annotator should use the term "radius" in their annotation of this fragmented entity because it is closer to the word "fracture" while retaining the meaning of the label.

This tool should not be used when annotating two different, isolated symptoms which are associated with the same AE. For example:

- “She was confused on the ward” – Is an appropriate token to select for the label *Delirium*
- “She was agitated on the ward” – Is also an appropriate token to select for the label *Delirium*
- “She was confused and agitated on the ward” - since both of these symptoms, in isolation, are sufficient to detect delirium, the tokens highlighted in this example must be annotated as two separate entities with the label *Delirium*. The discontinuous entity tool should not be used in this case.

B.3.3 Overlapping entities:

Medical texts also tend to contain overlapping entities which require the annotator to highlight the same element of text twice and this can be used in combination with the distant entities (Huang et al., 2023). For example:

- “When she fell, she broke her wrist and her hip” - This can be annotated with “broke” + “wrist” for the label *Wrist fracture* and “broke” + “hip” for the label *hip fracture*.

B.3.4 Supplementary instructions

Tokens The annotator should select tokens in their entirety, and never splitting them, even in the case of abbreviations and acronyms. For example:

- "She was admitted with a catheter-associated UTI and her catheter was replaced with antibiotic cover"
- "She was admitted with a CAUTI and her catheter was replaced with antibiotic cover" - These two examples demonstrate that although we do not need to include "catheter-associated" in our selection of tokens when labelling *Urinary tract infection*, we must include "CA" in "CAUTI" because the latter forms one, single token.

Only the necessary tokens should be highlighted. Pronouns, articles, and numbers should be excluded from the selection unless they are crucial to identifying AE. For example:

- "Xray confirms a fracture of the right wrist" - We do not want to include the side of the fracture (right or left) in our selection nor prepositions ("of") and articles ("the"), therefore only the terms "fracture" and "wrist" should be selected. This example makes use of the discontinuous entity tool.

We also recommend that annotators also avoid including punctuation in their selection wherever possible.

- "He had no melaena/haematochezia" - These two tokens can be isolated since they are separated by a slash (/)

We acknowledge that on occasions, the selection of punctuation cannot be avoided, for example:

- "The hip Xray confirmed a femoral #" - the hash sign "#" is often used by clinicians to mean "fracture"

Repetition: If an AE appears multiple times within a document, every instance of it should be annotated. For example:

- "She fell at home. This fall was likely caused by postural hypertension. She has been struggling with her gait since the fall because she lost confidence in her balance."

Number of fragments: When deciding to extract discontinuous entities, the annotator must restrict the use of the discontinuous tool to only two fragments. For example:

- "She fractured several bones in January including two ribs and the proximal end of her left femur.

Although our label *Hip fracture* specifically requires the annotator to extract fractures of the proximal end of the femur (as opposed to the distal end or mid-shaft fractures), we acknowledge that it would be impossible for them to extract the terms "femur", "fracture" and "proximal end" without either selecting a large number of unwanted tokens or using a third fragment with the discontinuous tool. For this reason, most entities will be defined in this document through two core elements which must be present in the selection of text. For example:

- *Lower respiratory tract infection* and *Urinary tract infection* must include a reference to the infection itself ("sepsis", "infective", "septic") and its source ("lung", "chest", "urinary", "bladder") although some words in isolation contain both pieces of information ("pneumonia", "pyelonephritis")
- All major osteoporotic fractures must include a term associated with the fracture itself ("broke", "fractured") and their location ("wrist", "radius", "humeral", "vertebra")

B.4 Annotation labels

The AE category has a set of 14 annotation labels which describe the information we are interested in capturing: hip fracture, proximal humeral fracture, wrist fracture, vertebral compression fracture, intracranial haemorrhage, upper gastrointestinal haemorrhage, lower gastrointestinal haemorrhage, urinary tract infection, lower respiratory tract infection, constipation and seizures. The following examples briefly demonstrate the annotation of these labels:

1. "Three falls in last yr" *(Falls)*
2. "On admission, she was acutely confused" *(Delirium)*
3. "Nursing staff has noted a decubitus ulcer on the right heel" *Pressure injury*
4. "Admitted with a left intertrochanteric neck of femur fracture" *(Hip fracture)*

5. "Distal radial fracture shown on the X-ray." *(Wrist fracture)*
6. "The pain in his shoulder is the result of a humeral fracture" *(Proximal humeral fracture)*
7. "She has several known osteoporotic vertebral fractures" *(Vertebral compression fracture)*
8. "Right parietal subdural haematoma visible on CT" *(Intracranial haemorrhage)*
9. "He went to see his GP the day before admission with abdominal pain and 3 episodes of melaena." *(Upper gastrointestinal haemorrhage)*
10. "She reported some PR bleed" *(Lower gastrointestinal Haemorrhage)*
11. "She was recently treated for pneumonia by her GP" *(Lower respiratory tract infection)*
12. "During her stay, she described a new burning sensation when passing urine. Ciprofloxacin was started for UTI." *(Urinary tract infection)*
13. "Patient had difficulties voiding despite laxatives. This was resolved with an enema." *(Constipation)*
14. "Found by the nurse with generalised tonic-clonic episode" *(Seizure)*

B.5 Annotation attributes

Annotation attributes are additional features or characteristics associated with the tag. For example, the tag *Person* could have the attribute age, gender or profession. For the tag *Organisation*, it could be the type of organisation, its number of employees, etc.

For this project, we use four main attributes: Negation, Diagnosis, in-hospital events and low confidence.

B.5.1 Negation:

Negation can be formulated explicitly with "no" or "not" but can also appear as acronyms (e.g. NAD, meaning nothing abnormal discovered) or be inferred from the text. The annotator should, when possible, not include the negation word(s) in the selection of text but simply highlight the words which reference the AE and add the negation attribute to the label. The following examples highlight how negation can be captured:

- "She slipped but did not fall." - This should be annotated *Falls* with the negation attribute.
- "CT scan found no evidence of intracranial bleeding." - This requires the label *Intracranial haemorrhage* and the negation attribute.

B.5.2 Low Confidence

The *Low confidence* should be selected when the annotator is uncertain of their choice. There are two main situations we recommend the use of this attribute:

- If there is insufficient information in the document (see examples below)
- If the annotator is faced with an unusual description of events which is, according to the writer of the letter, to be associated with AE.

In the second situation, the purpose of the *Low confidence* attribute is to provide some nuance to the final annotated document and indicate to the model that the text highlighted is unlikely to be associated with this AE, again, if in a different context. For example:

- "Her husband said that she had broken her leg about 6 months ago after a fall" - This should be annotated with the discontinuous entity tool and labeled *Hip fracture*. The *Low confidence* attribute should be added since we cannot be certain from this sentence alone that it was her proximal femur which the patient broke.
- "She was admitted with delirium, was confused in the ambulance and had visual hallucinations at home. Her daughter said she was describing children running around but they were by themselves in the house. The delirium resolved with rehydration." - Hallucinations are rarely present in delirium but it seems to be described by the author of the letter as such. It should therefore be annotated with the label *Delirium* and the *Low Confidence* attribute.

B.5.3 Diagnosis

This attribute allows the annotator to further qualify the annotated text when the AE is not directly present in the patient's admission, attendance or report. For this purpose, four tags are available:

- *History of*, when selected, qualifies AE which has happened in the past. This can be mentioned in the patient's past medical history or something which happened before the admission or attendance and is not related to the current clinical presentation.

- *Suspected*, qualifies when the writer of the clinical document suspects the patient has a GS or AE (note this is not the suspicion of the annotator - if the annotator is unsure about an AE to be annotated, nuance can be added with the attribute *Low confidence* instead (see previous paragraph)). This attribute can also be used in annotating screening tests which require further investigations to confirm a diagnosis.
- *Referral* qualifies when a patient is being referred to a different healthcare professional for the diagnosis of their AE.
- *Implicit mention* qualifies an AE which is mentioned in the text but does not apply to the patient. This includes descriptions of the patient's family history or diagnosis given to their spouse. It also should be used when the text details hypothetical situations ("if" statements) or risks of an AE happening (e.g. risk of falls).
- Finally, if the *Diagnosis* field is left empty, this signifies that the AE has been diagnosed, is present during the admission or attendance or that it is the reason for the admission or attendance.

The examples below illustrate each tag:

- "PMH: T2DM, Parkinson's disease, subdural haemorrhage." - In this example, the "subdural haemorrhage" has been described in the past medical history (PMH) section of the letter and should be annotated with the label *Intracranial haemorrhage* and the attribute *History of* for the *Diagnosis*.
- "His haemoglobin remained stable and was referred to the GI team to discuss outpatient investigations of this haematemesis"- In this example, the haematemesis should be annotated as *Upper gastrointestinal haemorrhage* with *referral* attribute.
- "CT Head showed acute on chronic subdural haematoma." - In this case, there is a historical feature (chronic) as well as a recent event (acute) of the intracranial bleed. The highlighted text should therefore be annotated twice with the same label (*Intracranial haemorrhage*), one with the attribute *History of* and one without any further attribute.
- "This patient was admitted following a GP visit who suspects she might have a wrist fracture." - Here the wrist fracture is *Suspected* and should be annotated with the matching *Diagnosis* attribute.
- "The patchy opacification in the left base may represent an infective process or pulmonary oedema." - This is an example of a radiology interpretation which maintains a level of doubt onto the final diagnosis. We recommend annotating this as one discontinuous entity of *Lower respiratory tract infection* with the *Suspected* attribute
- "The community occupational therapy team will review the patient's house after his discharge to evaluate and minimise the risk of falls." - This is a hypothetical mention of the GS *Falls* (with "risk of"). It should therefore be annotated as *Falls* with the attribute *Implicit mention*.
- "He was given laxatives for prophylaxis of constipation" - The term prophylaxis means the patient was given laxatives as a preventative measure, therefore, they did not experience constipation and it should be annotated with the attribute *Implicit mention*.
- "He was admitted with a fall and developed delirium during his admission due to a urinary tract infection." - All three of these diagnoses have happened in direct relation to the admission and therefore should be annotated without any additional *Diagnosis* attribute. Fall last night and was brought to the hospital today after a long lie." - Although the event happened in the past, it is directly related to the current admission therefore the fall should be annotated without any *diagnosis* attribute.

Some letters might have a template to help clinicians write quicker by simply answering propositions. If elements of the template which contain an AE are not applicable to the patient, they should be annotated with the *Implicit mention*. For example:

- "Type and duration of DVT prophylaxis (only for hip fracture): Not applicable" - *Implicit mention*

- "Type of Fall: / Site of the fracture: / Operation date:" - *Implicit mention*

Clinical documents often contain mentions of previous adverse drug reactions which name the drug and the exact adverse event which occurred (e.g. drug-induced delirium). These should be annotated with the attribute *History of* as shown in the example below:

- "Allergies and adverse reactions: Bisoprolol (dizziness), Metformin (constipation)" - This means the patient has experienced constipation because of the medication Metformin in the past, therefore it should be annotated with the attribute *History of*.

B.5.4 In-hospital event

This attribute allows the annotator to identify AE for which the onset took place in the hospital. For example, if a patient has a fall on the ward after surgery or if they were admitted with no symptom of delirium but developed acute confusion during their stay.

This attribute aims to capture the location of the event rather than the location of the diagnosis. This is important to keep in mind since some entities are diagnosed during the admission but their onset will have taken place before the patient came to the hospital. For example, if a patient had a fall two days prior and is admitted and treated for an infection. On day 3 they reported pain in their wrist and the X-ray of their arm shows a fracture but clearly, this adverse event is associated with their fall at home 5 days ago, therefore the attribute *In-hospital event* should not be selected. We present further examples in the following part:

- "New cough and wheeze during admission. Treated as hospital-acquired pneumonia" - In this case, "pneumonia" should be tagged with the label *Lower respiratory tract infection* and the attribute *In-hospital event* should be selected.

- "During her post-operative recovery, she described a new burning sensation when passing urine. CRP was raised, urine dip was positive for blood and leukocytes and Ciprofloxacin was started for UTI." - This should be tagged as with the label *Urinary tract infection* and the attribute *In-hospital event*.

B.6 More examples

We are providing in the following section a set of examples for each AE in order to emphasise the different expressions that can be found in free text and how they should be annotated. We are also providing two more examples on Brat in Figures 1 and 2.

Falls: Different expressions can be used for falls including:

- "Fell down steps at front door last week."

- "Dizzy when turns quickly, fell but only onto bed."

- "Found on floor by carer"

- "His wife mentioned that she is very worried he could fall when she is not watching him, the community occupational therapy team has planned to visit his house for a falls assessment"

This sentence includes "Fall" and should be annotated. It describes initially a hypothetical event and later an assessment of the risk factors of falls in the patient's house. Both of these cases should be annotated as *Implicit mention* because the geriatric syndrome has been described in the text but the event itself has not happened.

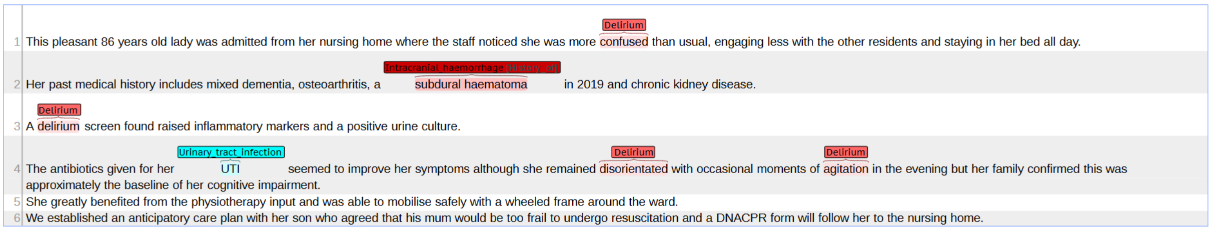
- "He started using a wheeled frame to avoid falling."

This is another example where the patient has not fallen. Just like a risk of falls, this example should be annotated with the attribute *Implicit mention*

Please note that falls should only be annotated in a geriatric context. For example, "fell off a wall while rock climbing" or "fell in horse riding accident" do not apply as these are not geriatric falls. Similarly, not all references to the noun or verb "fall" should be annotated. For example "falling sodium" or "patient moved house last fall" should be ignored.

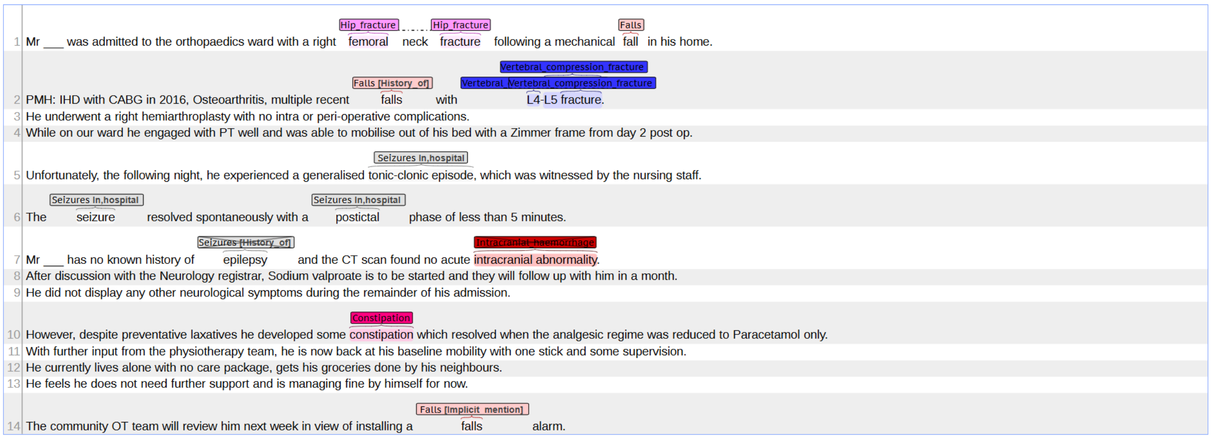
Delirium To annotate *Delirium*, the annotator will be looking for evidence of acute fluctuation or resolution of the impaired cognition. For example:

- "On admission, she was confused and agitated which settled to some extent with pain relief and rehydration."



This pleasant 86 years old lady was admitted from her nursing home where the staff noticed she was more confused than usual, engaging less with the other residents and staying in her bed all day.
Her past medical history includes mixed dementia, osteoarthritis, a subdural haematoma in 2019 and chronic kidney disease.
A delirium screen found raised inflammatory markers and a positive urine culture.
The antibiotics given for her UTI seemed to improve her symptoms although she remained disorientated with occasional moments of agitation in the evening but her family confirmed this was approximately the baseline of her cognitive impairment.
She greatly benefited from the physiotherapy input and was able to mobilise safely with a wheeled frame around the ward.
We established an anticipatory care plan with her son who agreed that his mum would be too frail to undergo resuscitation and a DNACPR form will follow her to the nursing home.

Figure 1: Example 3 on Brat



Mr ___ was admitted to the orthopaedics ward with a right femoral neck fracture following a mechanical fall in his home.
PMH: IHD with CABG in 2016, Osteoarthritis, multiple recent falls with L4-L5 fracture.
He underwent a right hemiarthroplasty with no intra or peri-operative complications.
While on our ward he engaged with PT well and was able to mobilise out of his bed with a Zimmer frame from day 2 post op.
Unfortunately, the following night, he experienced a generalised tonic-clonic episode, which was witnessed by the nursing staff.
The seizure resolved spontaneously with a postictal phase of less than 5 minutes.
Mr ___ has no known history of epilepsy and the CT scan found no acute intracranial abnormality.
After discussion with the Neurology registrar, Sodium valproate is to be started and they will follow up with him in a month.
He did not display any other neurological symptoms during the remainder of his admission.
However, despite preventative laxatives he developed some constipation which resolved when the analgesic regime was reduced to Paracetamol only.
With further input from the physiotherapy team, he is now back at his baseline mobility with one stick and some supervision.
He currently lives alone with no care package, gets his groceries done by his neighbours.
He feels he does not need further support and is managing fine by himself for now.
The community OT team will review him next week in view of installing a falls alarm.

Figure 2: Example 4 on Brat

This should be annotated as *Delirium* even though it was not explicitly stated because we can see that the cognitive impairment has been resolved (albeit partially). Note that a precipitant can be identified in some cases (e.g. clues of bacterial infection alongside the cognitive impairment and improved cognition under antibiotics) and support the annotator's confidence. However, it is not a requirement for 10% of delirium diagnoses do not have a clear precipitant (MacLullich et al., 2009).

If the text suggests a cause for the patient's altered cognition which is not delirium, it should be ignored by the annotator, for example:

- "Patient was forgetful, disoriented: this was believed to be the result of metastatic disease to the brain."

Pressure injury: Only ulceration or damage to the skin as a result of prolonged pressure should be included in this category. Other injuries such as venous ulcers and arterial ulcers should not be highlighted. If the type of injury is not explicitly named, its location[12] is a strong marker, as shown in these examples:

- "Pressure sore right buttock, down to muscle"
- "Known diabetic foot, skin break the right metatarsal head. Stage 3 bedsore on left shoulder"
- "Hot spot on left heel, deep ulcer right ankle"

An ulcer on the ankle is an unlikely location for a pressure injury and is presumably the result of venous insufficiency. It should therefore be ignored in the annotation.

Hip fractures: As defined in Appendix 2, this label includes all proximal femoral fractures except isolated greater trochanter fractures and periprosthetic fractures. Different expressions can be found in free text to refer to hip fractures, including:

- "Fracture of the left neck of femur" - In this example we want to include the term "fracture" as well as "femur" so we use a discontinuous entity to highlight both these elements as *Hip fracture*.
- "He broke his right hip last year" - This is a discontinuous entity and should include the *History of* attribute
- "Right intracapsular neck of femur fracture" - This a straightforward example which does

[12] https://www.cancerresearchuk.org/about-cancer/coping/physically/skin-problems/pressure-sores/causes-and-prevention

not require the use of discontinuous entities.

- "Pelvis and R Hip X-ray request: Ms___ had a fall last night and landed on her right side. She was unable to stand and on examination, her left leg was shortened and externally rotated. NOF#?" - This is a radiology request (NOF means Neck of femur and # means fracture) therefore this selection should be annotated with the *Suspected* attribute.
- "Isolated fracture of the greater trochanter of the left femur." - This element should not be annotated. As described in Appendix 2, our definition of hip fractures excludes isolated greater trochanter fractures.
- "R Hip X-ray: Some arthritic changes with narrowing of the joint space and osteophytes. No visible fracture." - This should be annotated with the discontinued entity *Hip fracture* and the *Negation* attribute.
- "This gentleman had a fall in his house, unable to weight bear since. ?Fracture of the left hip. Radiologists report: Left dynamic hip screw in position. No fracture observed." - We exclude in our annotation of hip fractures any periprosthetic fracture. Hence, the mention of a dynamic hip screw means this section of text should not be annotated.

Wrist fractures: They are defined as fractures of the distal end of the radius, the distal end of the ulna or both. Different expressions can be found in free text to reference wrist fractures, including:

- "Colles fracture on the right side" - A Colles fracture is a specific fracture of the distal radius with dorsal angulation and displacement. It needs to be included in the selection because it locates the fracture on the wrist.
- "The X-ray showed an ulnar styloid fracture" - This is a discontinuous entity
- "She fell on her wrist and the X-ray showed a fracture of the 5th metacarpal" - This should not be annotated as we only want to identify fractures of the distal ulna and radius.
- "Wrist X-ray report: Identifiable radial fracture with posterior displacement."
- "The wrist X-ray performed in A&E showed no fracture" - This is a discontinuous entity with the *Negation* attribute
- "Imaging shows a right impacted and comminuted distal radial fracture with associated avulsion fracture of the ulnar styloid process." - *Radial fracture* is one continuous entity. And *fracture* and *ulnar* are a separate discontinuous entity.

Proximal humeral fractures: Different expressions can be found in free text to reference proximal humeral fractures, including:

- "X-ray confirmed a left humeral neck fracture"- This is a discontinuous entity
- "Her son reports that she had a broken shoulder after a fall 6 months ago" - This is a discontinuous entity and should be annotated with the attributes *History of*

Note that since the shoulder contains multiple bones and the terms "broken shoulder" are not specific to the proximal humerus, if the rest of the text does not specify its location any further, then this expression should be annotated with *Low confidence*.

- "CXR report: Both fields are clear, evidence of healed right neck of humerus fracture."

Here the X-ray was performed to evaluate the patient's chest and not their shoulder or upper arm. Nonetheless, "humerus fracture" should be annotated along with the attribute *History of*. This example will be detailed and annotated further in the *Lower respiratory tract infection* paragraph.

Vertebral compression fractures: The majority of vertebral fractures are the result of osteoporosis and unless the text mentions a risk of bony metastases from cancer or a high-velocity trauma, all occurrences of a fracture of vertebral body should be annotated with this entity. Fractures of the vertebral arch should be ignored. Most osteoporotic vertebral fractures will occur in the mid-thoracic and thoracolumbar regions below the T4 level(Griffith, 2015) therefore any fractured vertebra above T4 level should not be annotated unless identified in the text as a result of osteoporosis. We recommend whenever possible to include the position of the fractured vertebra (T/L + number) in the selection of text. Different expressions can be found in free text to reference vertebral compression fractures, including:

- "PMH: T2DM, L2 fracture" - This also requires the *History of* attribute.

- "Osteoporotic fracture of L4" - In this example, the highlighted text is one discontinuous entity.
- "Crush fractures of T12 to L3 - In this example, we recommend using discontinuous entities and overlapping the selection of text so that two entities are extracted from this sentence: "fracture + T12" and "fracture + L3"
- "His wife mentioned that he was diagnosed with two broken vertebrae in his lower back a month ago" - This would also require the *History of* attribute.
- "X-ray revealed fractures of T9 to 11" - In this example, we can clearly annotate "fractures + T9" but annotating "fractures + 11" would not make sense because "T" is crucial to the site of the fracture. Therefore, our second selection should include "T9 to 11" to the fragment and this examples should be annotated as one discontinuous entity.
- "Lumbar Spine Xray: Image compared to December 2012. Noted vertebral fractures of L4 and L5 present on previous images. Evidence of new fracture of L4 with further loss of height" - In this example again we recommend annotating "vertebral fracture" and the use of discontinuous and overlapping entities for "fractures + L4" and "fractures + L5" alongside the *History of* attribute. As for the last sentence of the example, we recommend selecting the text as one discontinuous entity and with no attribute.

This entity should also be used with the negation attribute whenever a Thoracic or Lumbar X-ray shows no fracture. For example:

- "Lumbar Spine XR report: Multi level degenerative changes throughout the lumbar spine. No fracture detected." - This is one discontinuous entity with the *Negation* attribute.
- "She was admitted with back pain following a fall. X-ray was normal and she progressed well with PT." - This is one discontinuous entity with the *Negation* attribute.

Intracranial haemorrhage: Intracranial haemorrhages are diagnosed on imaging of the head, most frequently used is the CT scan. Most CT scans of the head performed after a fall in the elderly or during an episode of delirium will be looking for an intracranial bleed. We therefore encourage the annotation of any normal CT scan result with this entity and the *Negation* attribute. Different expressions can be found in free text to reference intracranial haemorrhage, such as:

- "History of haemorrhagic stroke at 52" - This should be annotated with the *History of* attribute.
- "T2-FLAIR shows hyperdensity in keeping with subarchnoid haemorrhage of left temporal region"
- "Head CT found no mass or bleed" - Both of these elements should be annotated as one discontinuous entity and with the *Negation* attribute.
- "CT head shows acute subdural haematoma" - The term "acute" should not be included in the selection because it does not help us identify the AE but simply indicates to us that this is a new or recent diagnosis.
- "CT head shows chronic subdural haematoma" - The term "chronic" should not be included in the selection however the attribute *History of* should be added to the annotation.
- "CT Head shows acute on chronic subdural haematoma" - This selection of text should be highlighted twice, both with the entity *Intracranial haemorrhage*, one with no *Diagnosis* attribute and the other with the attribute *History of*.
- "CT Head: is unremarkable" - Both of these elements should be annotated as one discontinuous entity and with the *Negation* attribute.
- "CT Head: No intra-axial or extra-axial haemorrhage. No infarct visualised. Conclusion: No intracranial abnormality." - This description of a CT scan with no intracranial haemorrhage requires three separate entities to be annotated. The first one is a discontinuous entity "intra-axial + haemorrhage", the second one is "extra-axial haemorrhage" (continuous) and the last one is "intracranial abnormality". Each of these elements of text should be annotated as *Intracranial haemorrhage* with the *Negation* attribute.
- "CT scan of head. Noted history of subdu-

ral haematoma from a traumatic head injury six months ago, still present and stable on imaging. No new intracranial haemorrhage" - The first element should be annotated with the attribute *History of* and the second selection of text being preceded by "no new" reminds us that the patient had an intracranial bleed in the past while confirming there is no haemorrhage at present. Therefore it should be annotated once with the *History of* attribute and a second time (selecting the exact same words) with the *Negation* attribute.

- "CT shows an extensive cerebellar haemorrhage"

Upper gastrointestinal haemorrhage: The context around the portion of text we want to annotate might be useful to differentiate upper GI bleed from lower GI bleed:

- "While in A&E, she had coffee-ground emesis"
- "He underwent emergency gastroscopy after 5 episodes of haematemesis which found ruptured oesophageal varices." - These are two distinct entities to be annotated with the same label
- "EOGD found evidence of gastritis and no active bleeding" - Discontinuous entity with the *Negation* attribute
- "During the admission she had two episodes of melaena"

Lower gastrointestinal haemorrhage: Different expressions can be used for lower GI haemorrhage, including:

- "I reviewed him at the surgery today for PR bleed"
- "85-year-old male with a history of diverticulosis presented to the acute medical unit for new breathlessness on exertion and melaena. Low HB was supplemented with 2 units of RBC, OGD was unremarkable, likely bleed from diverticulosis." - Melaena is usually a clinical sign of upper gastrointestinal haemorrhage but here the context indicates this is more likely a lower GI haemorrhage.
- "Reported diarrhoea with mucus and blood" - Discontinuous entity
- "The patient may notice traces of blood in their stools in the weeks following the surgery" - This is to be annotated as one discontinuous entity and with the *Implicit mention* attribute

Despite their potential for detecting lower gastrointestinal haemorrhage, we recommend against the annotation of any bowel cancer screening tool such as faecal occult blood tests.

Please also note that clinicians might simply mention "gastro-intestinal haemorrhage" in the patient's records with no further specification. In this case, the annotator should highlight the expression with both entities *Upper gastrointestinal haemorrhage* and *Lower gastrointestinal haemorrhage* as well as the *Low confidence* attribute.

Urinary tract infection: For this entity, the annotator needs to find a mention of both the infection and its location on the urinary tract. Symptoms (frequency, burning sensation...), lab results (positive dip, cultures...) or treatment (antibiotics) should not be annotated. This is reflected in the examples below:

- "Reason for admission: Pyelonephritis"
- "Admitted with delirium. Urinalysis was positive and he was treated for a UTI with good response."
- "Past medical history: Alzheimer's disease, GORD, Multiple CAUTI" - The attribute *History of* should be selected (CAUTI means catheter-associated urinary tract infection)
- "During this admission, he became more confused and his inflammatory markers raised. Urinalysis was positive, urine culture was sampled and he started Nitrofurantoin based on previous sensitivities. The infection settled after two days and he was able to be discharged back home with no further delay." - This should be annotated as one discontinuous entity and with the *In-hospital event* attribute.

Please note that a positive urine culture or urinalysis is not sufficient to confirm a urinary tract infection. Instead, the annotators should look for terms such as "infection" or "sepsis" and "urine" or "urinary tract", using the discontinuous entity tool when necessary.

Lower respiratory tract infection Healthcare professionals often specify whether a lower respiratory tract infection was acquired in the community or in the hospital as this will guide the antibiotics they prescribe. If this is not explicit, the context around the annotated text can help guide the decision of the annotator to add the *In-hospital event* attribute. Similarly to the entity *Urinary tract infection*, the annotator needs to find a mention of both the infection (sepsis, septic, infective...) and its location on the lower respiratory tract (lung, chest, bronchial, pulmonary...). Symptoms (cough, shortness of breath...) , lab results (sputum cultures) and treatment (antibiotics) should be ignored. See the examples below for more details:

- "New cough and wheeze during admission. Treated as hospital-acquired pneumonia" - This should be annotated with the attribute *In-hospital event*
- "Treated for ventilator-associated pneumonia" - This should be annotated with the attribute *In-hospital event*
- "Admitted from a care home for breathlessness, fever and chills. CXR was consistent with aspiration pneumonia." - This should be annotated without the attribute *In-hospital event*.
- "She received treatment 3 months ago for CAP" - This should be annotated with the *History of* attribute. "CAP" refers to "community-acquired pneumonia" so this should be annotated without the attribute *In-hospital event*.
- "She received antibiotics post operatively for a HAP" - Here "HAP" means hospital-acquired pneumonia and should be annotated with the *In-hospital event* attribute.
- "CXR report: New consolidation in the left lower lobe"
- "Reason for admission: Infective exacerbation of COPD" - Annotated as one discontinuous entity
- "Three episodes of vomiting in A&E. Now coughing. CXR requested for ?aspiration. - In this context, "aspiration" refers to "aspiration pneumonia" which is a lower respiratory tract infection following the inhalation of food, liquid or vomit. The annotation of the word "aspiration" on its own is recommended in this situation, alongside the *Suspected* attribute
- "CXR confirms there is an infective process in her right lung" - One discontinuous entity

Since tuberculosis can manifest in various organs and systems within the body, we recommend that if the annotator selects it, they also add a location in the chest/lungs. For example

- "CT scan shows a cavity in the apex of the left lung. This appears to be scarring from previous TB as demonstrated by previous imaging" - This would be annotated using the discontinuous entity tool and the attribute *History of* should be added

The term "consolidation" when described in the report of an imaging of the chest almost always refer to an infection. When it does not, the radiologist or clinician will usually specify the differential diagnosis (e.g. "consolidation consistent with progression of the lung tumour"). Therefore, based on the context of the document, the annotators are encouraged to annotate this terms as *Lower respiratory tract infection* if it appear to match this diagnosis.

- "CXR showed a new patchy consolidation of the right lower lobe. Antibiotics started for pneumonia" - These are two separate entities annotated with the same label.

Most chest X-rays requested during the admission of an older adult will be looking for signs of an infection. Any normal chest X-ray will almost always exclude a lower respiratory tract infection. We therefore encourage annotators to highlight normal chest X-ray results with this entity and the negation attribute, as shown below:

- "CXR is normal" - To be annotated as one discontinuous entity with the *negation* attribute.
- "CXR report: Both lung fields are clear, evidence of healed right neck of humerus fracture." - For the *Lower respiratory tract infection* entity, we want to annotate the above sections of text as one discontinuous entity with the *Negation* attribute. This example was also reviewed in the *Proximal humeral fractures* paragraph.

Note that the acronym "HCAP", or "HealthCare Associated Pneumonia" is often used by clinicians in their discharge letters. These describe a chest infection developed in the context of healthcare interventions such as an admission to the hospital or a stay in rehabilitation facility or long term care facility (e.g nursing home or care home). Our *In-hospital event* attribute however, should only be used to identify GS or AE developed in hospital (acute or sub-acute care). Therefore a pneumonia diagnosed in the patient's records as "HCAP" simply because the patient was admitted from a nursing home or a care home should not be annotated with this attribute. On the other hand, an "HCAP" diagnosed part-way through the admission (not present when the patient initially entered the hospital) should be annotated as a *Lower respiratory tract infection* with the *In-hospital event* attribute.

Constipation: Different expressions can be found in free text to refer to constipation, such as:

- "Bowels not opened in 5 days"
- "No recorded BM in 4 days"
- "She reported a slower transit than usual"
- "Adverse drug reactions: Amoxicillin (rash), Co-codamol (vomiting), Metformin (constipation)" - This should be annotated with the attribute *History of*.
- "Discharge medications: Morphine sulphate 2mg every 4 hours as required, Senna 7.5 mg up to twice a day as required for constipation" - In this context, the prescription of a laxative "as required" implies the patient might not be experiencing constipation. We recommend annotating any mention of the entity *Constipation* with the attribute *Implicit mention* when it appears in the medication list and described as "as required". This avoids extracting constipation when a laxative is given preventatively or absentmindedly left in a repeated prescription. If the patient reported constipation during their admission/presentation, the rest of the document would likely contain a mention of it and it would then be annotated without the attribute *Implicit mention*.

Seizures: Different expressions can be found in free text in reference to seizures, including:

- "Postictal phase inferior to 5min"
- "First grand mal last month"
- "Family reported she had a fit in the waiting room"

C More statistics about the annotated dataset

Figure 3 illustrates the distribution of fine-grained entities in the train, dev and test corpora.

D Detailed results

D.1 Best and worst examples for the entity level results (IAA)

Table 2 illustrates the five best and worst results (entity-level) related the agreement among the two annotator for both the pilot and the test corpora.

D.2 Best and worst examples for the document level results (IAA)

Table 3 illustrates the five best and worst results (document-level) related the agreement among the two annotator for both the pilot and the test corpora.

D.3 Detailed results (entity level)

Figures 4, 5 and 6 illustrates the detailed precision, recall and f1-scores obtained for each entity in the dataset. We also highlight the frequency of each entity within the datset. For figure 4, we regroup all the entity having scores equal to 0 using the label "Other_tags". In Other_tags, we find rare entities such as **Lower_respiratory_tract_infection+In-hospital_event+Urinary_tract_infection+In-hospital_even** or **Seizures+Diagnosis+History_of+Negation**.

E More examples for the error analysis part

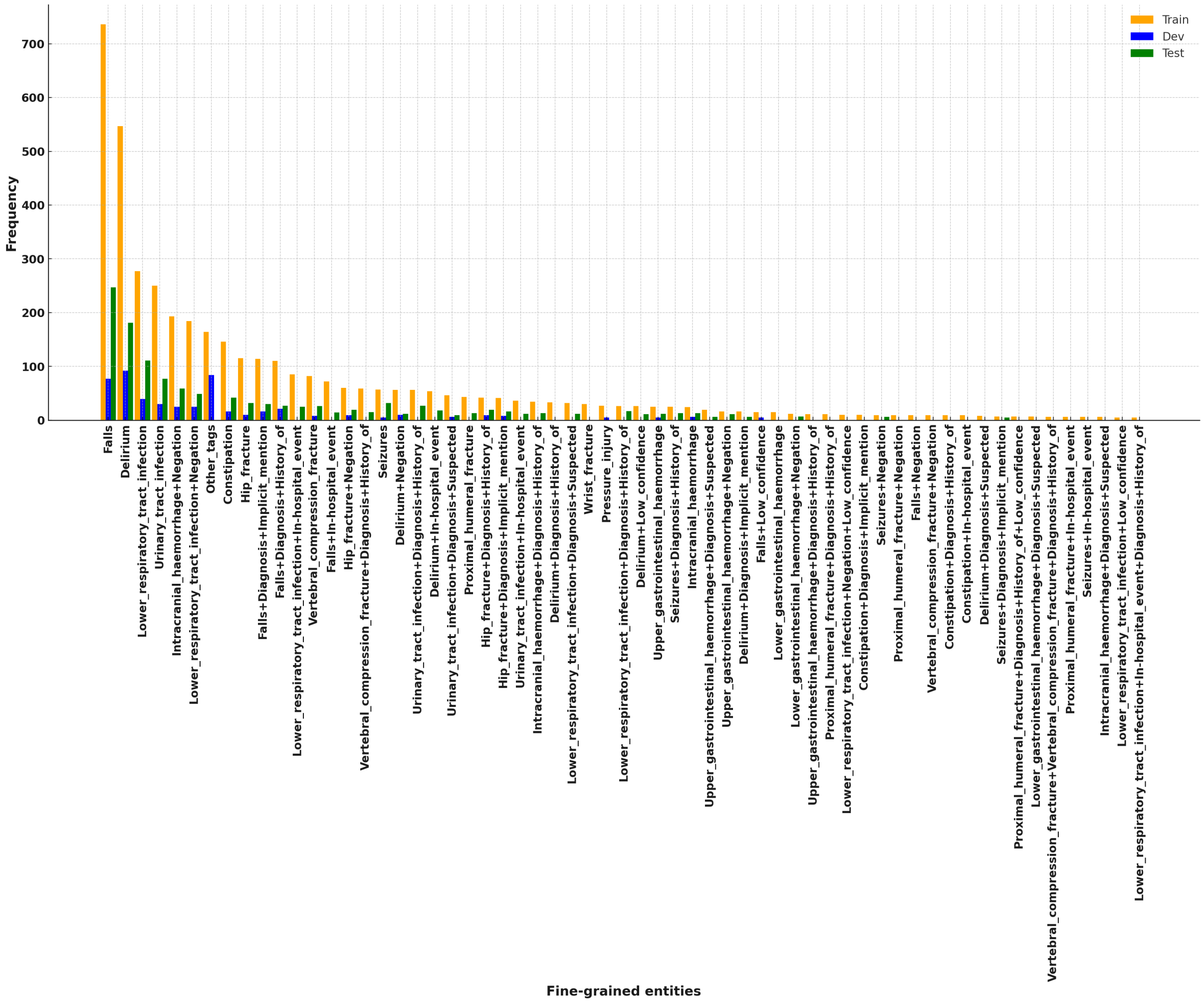


Figure 3: The distribution of fine-grained entities in the train, dev and test datasets

Table 2: IAA between the two annotators (entity level) using bratiaa

		Best entities		Worst entities	
Corpus	Annotation type	Entity	f1-score	Entity	f1-score
Pilot	Fine-grained	Delirium+Diagnosis+Suspected	1.000	Vertebral_compression_fracture+Diagnosis+History_of	0.000
		Hip_fracture+Diagnosis+History_of	1.000	Constipation+Diagnosis+History_of	0.000
		Seizures+Diagnosis+History_of	1.000	Intracranial_haemorrhage+Negation	0.000
		Upper_gastrointestinal_haemorrhage+Diagnosis+History_of	1.000	Lower_respiratory_tract_infection+Diagnosis+History_of	0.000
		Urinary_tract_infection+Diagnosis+History_of	1.000	Proximal_humeral_fracture+Negation	0.000
	Coarse-grained	Pressure_injury	1.000	Wrist_fracture	0.000
		Falls	0.912	VIntracranial_haemorrhage	0.000
		Upper_gastrointestinal_haemorrhage	0.875	Vertebral_compression_fracture	0.111
		Seizures	0.769	Proximal_humeral_fracture	0.222
		Urinary_tract_infection	0.760	Lower_respiratory_tract_infection	0.596
	Coarse-grained+negation	Pressure_injury	1.000	Wrist_fracture	0.000
		Upper_gastrointestinal_haemorrhage	0.933	Urinary_tract_infection+Negation	0.000
		Falls	0.912	Upper_gastrointestinal_haemorrhage+Negation	0.000
		Intracranial_haemorrhage+Negation	0.750	Proximal_humeral_fracture+Negation	0.000
		Delirium+Negation	0.667	Lower_respiratory_tract_infection+Negation	0.000
Test	Fine-grained	Lower_gastrointestinal_haemorrhage+Diagnosis+History_of	1.000	Seizures+Diagnosis+Referral	0.000
		Lower_gastrointestinal_haemorrhage+Diagnosis+Suspected	1.000	Wrist_fracture+In-hospital_event	0.000
		Intracranial_haemorrhage+Diagnosis+History_of	1.000	Vertebral_compression_fracture+Negation	0.000
		Hip_fracture+In-hospital_event	1.000	Urinary_tract_infection+Negation	0.000
		Falls+Negation+Diagnosis+History_of	1.000	Upper_gastrointestinal_haemorrhage+Diagnosis+Suspected+In-hospital_event	0.000
	Coarse-grained	Wrist_fracture	1.000	Vertebral_compression_fracture	0.488
		Urinary_tract_infection	0.927	Pressure_injury	0.500
		Falls	0.909	Intracranial_haemorrhage	0.523
		Seizures	0.885	Hip_fracture	0.569
		Constipation	0.835	Upper_gastrointestinal_haemorrhage	0.697
	Coarse-grained+negation	Proximal_humeral_fracture+Negation	1.000	Vertebral_compression_fracture+Negation	0.000
		Constipation+Negation	1.000	Urinary_tract_infection+Negation	0.000
		Lower_gastrointestinal_haemorrhage	1.000	Hip_fracture+Negation	0.291
		Seizures+Negation	0.833	Intracranial_haemorrhage+Negation	0.350
		Falls+Negation	0.800	Upper_gastrointestinal_haemorrhage+Negation	0.000

Table 3: IAA between the two annotators (document level)

		Best entities			Worst entities		
Corpus	Annotation type	Entity	f1-score	Cohen's kappa	Entity	f1-score	Cohen's kappa
Pilot	Fine-grained	Falls+Implicit_mention	1.000	1.000	Falls+In-hospital_event	0.000	0.000
		Delirium+Suspected	1.000	1.000	Upper_gastrointestinal_haemorrhage+Suspected	0.000	0.000
		Hip_fracture+History_of	1.000	1.000	Lower_respiratory_tract_infection+In-hospital_event	0.000	0.000
		Proximal_humeral_fracture+History_of	1.000	1.000	Urinary_tract_infection+Suspected	0.333	0.296
		Falls+Diagnosis	0.800	0.779	Lower_respiratory_tract_infection+Suspected	0.500	0.485
	Coarse-grained	Pressure_injury	1.000	1.000	Proximal_humeral_fracture	0.800	0.789
		Intracranial_haemorrhage	1.000	1.000	Delirium	0.85	0.792
		Upper_gastrointestinal_haemorrhage	1.000	1.000	Hip_fracture	0.860	0.846
		Constipation	1.000	1.000	Urinary_tract_infection	0.875	0.816
		Seizures	1.000	1.000	Lower_respiratory_tract_infection	0.96	0.946
	Coarse-grained+negation	Intracranial_haemorrhage+Negation	1.000	1.000	Urinary_tract_infection+Negation:	0.000	0.000
		Delirium+Negation	1.000	1.000	Upper_gastrointestinal_haemorrhage+Negation	0.000	0.000
		Constipation	1.000	1.000	Proximal_humeral_fracture+Negation	0.000	0.000
		Seizures	1.000	1.000	Delirium	0.8333	0.781
		Proximal_humeral_fracture	1.000	1.00	Lower_respiratory_tract_infection+B-Negation	0.857	0.847
Test	Fine-grained	Lower_gastrointestinal_haemorrhage+Suspected	1.000	1.000	Intracranial_haemorrhage+Suspected	0.000	0.000
		Proximal_humeral_fracture+Negation	1.000	1.000	Vertebral_compression_fracture+Negation	0.000	0.000
		Intracranial_haemorrhage+History_of	1.000	1.000	Hip_fracture+In-hospital_event	0.000	0.000
		Hip_fracture+History_of	0.963	0.929	Upper_gastrointestinal_haemorrhage+In-hospital_event	0.000	0.000
		Intracranial_haemorrhage+Diagnosis	0.923	0.857	Falls+Referral	0.000	0.000
	Coarse-grained	Wrist_fracture	1.000	1.000	Pressure_injury	0.5	0.497
		Falls	0.989	0.983	Proximal_humeral_fracture	0.769	0.765
		Intracranial_haemorrhage	0.966	0.961	Lower_gastrointestinal_haemorrhage	0.824	0.820
		Urinary_tract_infection	0.956	0.945	Constipation	0.836	0.821
		Lower_respiratory_tract_infection	0.944	0.920	Hip_fracture	0.841	0.817
	Coarse-grained+negation	Constipation+Negation	1.000	1.000	Urinary_tract_infection+Negation	0.000	0.000
		Wrist_fracture	1.000	1.000	Proximal_humeral_fracture+Negation	0.000	0.000
		Falls	0.989	0.984	Vertebral_compression_fracture+Negation	0.333	0.329
		Intracranial_haemorrhage+Negation	0.978	0.976	Pressure_injury	0.5	0.976
		Urinary_tract_infection	0.968	0.961	Upper_gastrointestinal_haemorrhage+Negation	0.615	0.610

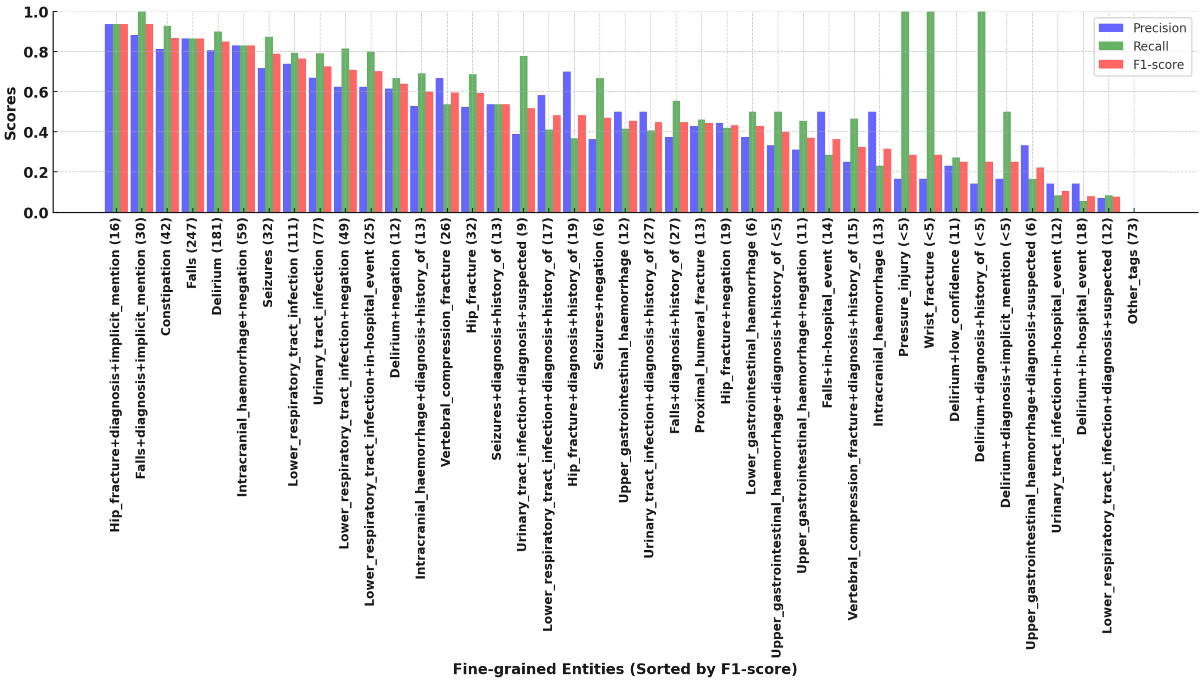


Figure 4: The results (precision, recall, F1-score) related to each fine_grained entity within the test corpus (obtained with BioBERT)

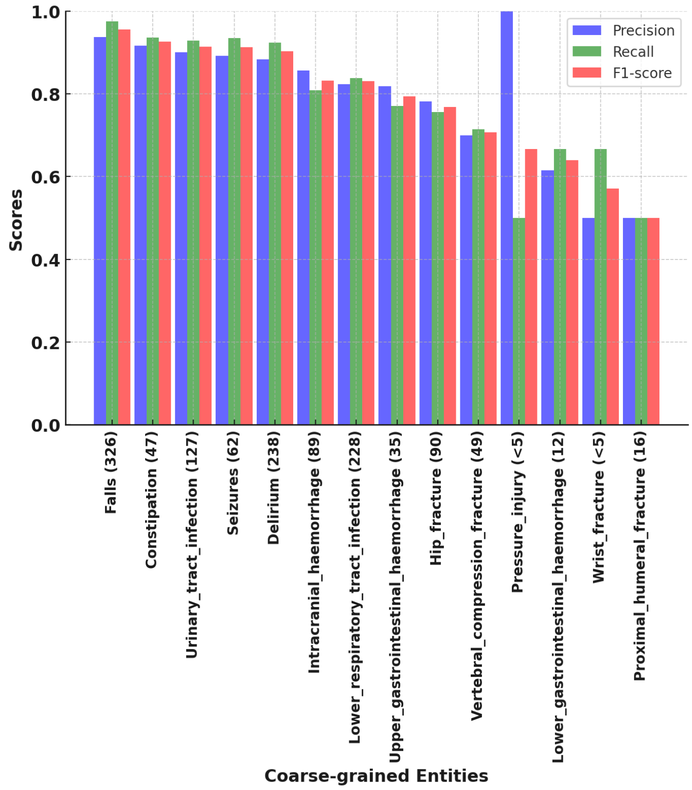


Figure 5: The results (precision, recall, F1-score) related to each coarse-grained entity within the test corpus (obtained with BERT_cased)

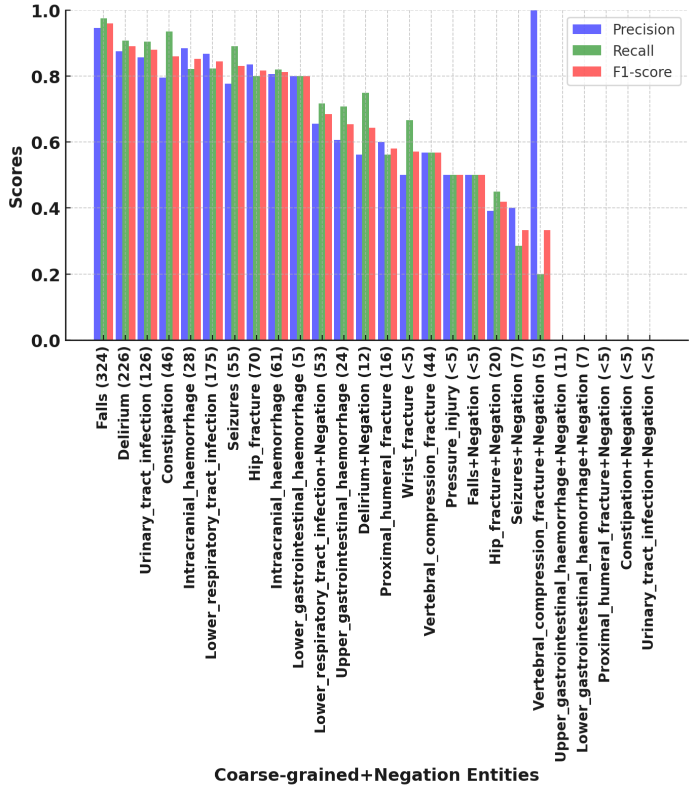
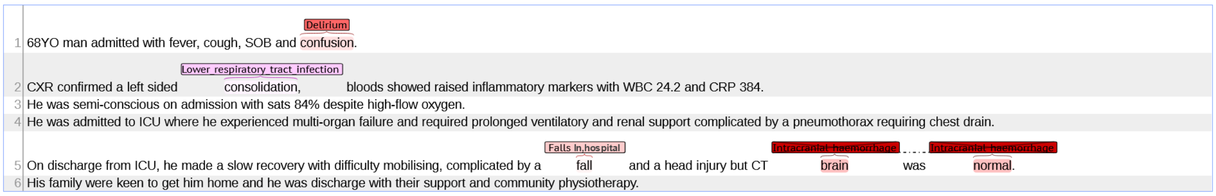
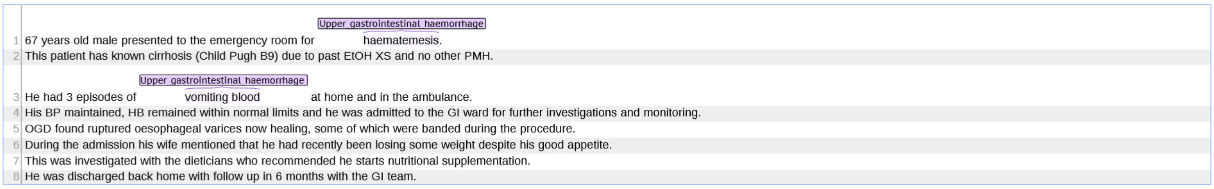


Figure 6: The results (precision, recall, F1-score) related to each coarse-grained+negation entity within the test corpus (obtained with BioBERT)



1 68YO man admitted with fever, cough, SOB and confusion.
2 CXR confirmed a left sided consolidation, bloods showed raised inflammatory markers with WBC 24.2 and CRP 384.
3 He was semi-conscious on admission with sats 84% despite high-flow oxygen.
4 He was admitted to ICU where he experienced multi-organ failure and required prolonged ventilatory and renal support complicated by a pneumothorax requiring chest drain.
5 On discharge from ICU, he made a slow recovery with difficulty mobilising, complicated by a fall and a head injury but CT brain was normal.
6 His family were keen to get him home and he was discharge with their support and community physiotherapy.

Figure 7: Example 2 on Brat


Upper_gastrointestinal_haemorrhage
Upper_gastrointestinal_haemorrhage

1 67 years old male presented to the emergency room for haematemesis.
2 This patient has known cirrhosis (Child Pugh B9) due to past EtOH XS and no other PMH.
3 He had 3 episodes of vomiting blood at home and in the ambulance.
4 His BP maintained, HB remained within normal limits and he was admitted to the GI ward for further investigations and monitoring.
5 OGD found ruptured oesophageal varices now healing, some of which were banded during the procedure.
6 During the admission his wife mentioned that he had recently been losing some weight despite his good appetite.
7 This was investigated with the dieticians who recommended he starts nutritional supplementation.
8 He was discharged back home with follow up in 6 months with the GI team.

Figure 8: Example 3 on Brat

Table 4: Some examples

Example	Annotation type	Entity level	Document level
Example 2 (Figure 7)	Fine-grained	**Span[10:11]**: confusion/Delirium(0.9996)	Delirium
		Span[17:18]: consolidation/Lower_respiratory_tract_infection(1.0)	Lower_respiratory_tract_infection
		Span[48:49]: fall/Falls(1.0)	Falls
		Span[55:58]: brain was normal/Intracranial_haemorrhage+Negation(0.9992)	Intracranial_haemorrhage+Negation
	Coarse-grained	**Span[10:11]**: confusion/Delirium(1.0)	Delirium
		Span[17:18]: consolidation/Lower_respiratory_tract_infection(1.0)	Lower_respiratory_tract_infection
		Span[48:49]: fall/Falls(1.0)	Falls
		Span[55:58]: brain was normal/Intracranial_haemorrhage(0.9998)	Intracranial_haemorrhage
	Coarse-grained+negation	**Span[10:11]**: confusion/Delirium(0.9698)	Delirium
		Span[17:18]: consolidation/Lower_respiratory_tract_infection(1.0)	Lower_respiratory_tract_infection
		Span[48:49]: fall/Falls(1.0)	Falls
		Span[55:58]: brain was normal/Intracranial_haemorrhage+Negation(0.9979)	Intracranial_haemorrhage+Negation
Example 3 (Figure 8)	Fine-grained	**Span[10:11]**: haematemesis/Upper_gastrointestinal_haemorrhage+ Diagnosis+History_of(0.6954)	Upper_gastrointestinal_haemorrhage
		Span[36:37]: vomiting/Urinary_tract_infection+In-hospital_event+ Lower_respiratory_tract_infection+In-hospital_event(0.0399)	
		Span[37:38]: blood/Lower_respiratory_tract_infection+ Diagnosis+Suspected+Lower_respiratory_tract_infection+Negation(0.0848)	
	coarse-grained	**Span[10:11]**: haematemesis/Upper_gastrointestinal_haemorrhage(0.9994)	Upper_gastrointestinal_haemorrhage
	coarse-grained+Negation	**Span[10:11]**: haematemesis/Upper_gastrointestinal_haemorrhage(0.9969)	Upper_gastrointestinal_haemorrhage