

A Human-Annotated Reference Corpus for Medication Information Extraction from MIMIC-IV Discharge Summaries

Provia Kadusabe¹, Abhishek Kaushik², Fiona Lawless¹

¹Dundalk Institute of Technology, Ireland

²Maynooth University, Ireland

{provia.kadusabe, fiona.lawless}@dkit.ie

abhishek.kaushik@mu.ie

Abstract

Medication information extraction from clinical notes supports medication reconciliation, pharmacovigilance, and clinical decision support, helping reduce medication errors and preventable patient harm. However, the linguistic variability and contextual dependence make medication mentions and their associated attributes difficult to identify consistently. Human-annotated reference corpora are therefore important for system development and rigorous evaluation, yet remain scarce. MIMIC-IV-Note provides a large-scale collection of discharge summaries but lacks a human-annotated, medication-specific reference corpus. We present a corpus of 305 MIMIC-IV-Note discharge summaries annotated by four annotators using a written guideline and adjudicated into a single reference version. The corpus covers seven entity types, a three-valued medication-status attribute, and five medication-attribute relation types. Pairwise inter-annotator agreement reached an exact-span F-measure of 0.896 and a character-level F-measure of 0.927. Medication mentions are split between medication-list sections (55.4%) and narrative sections (44.6%), with attribute information concentrated in the list sections (2.86 versus 0.59 attribute types per mention). A supervised entity-recognition baseline achieved a micro-averaged exact-span F-measure of 0.945 against the adjudicated reference on the held-out test set. Error analysis showed that most remaining errors involved span-boundary disagreements, particularly for the reason and duration attributes. We plan to release the corpus through PhysioNet to support reproducible development and evaluation of medication information extraction systems on MIMIC-IV-Note and to provide a curated human reference for assessing automatically generated medication labels.

Keywords: medication information extraction; clinical NLP; MIMIC-IV-Note; named entity recognition; relation extraction

1 Introduction

Medication information is central to clinical documentation. It is distributed across structured electronic health record fields and unstructured clinical notes. Structured medication lists and orders provide an accessible record of prescribed medications, but they do not necessarily capture the complete medication history. Clinical notes may contain more information such as medication changes that are not represented in structured fields, including medication initiation, discontinuation, and dose modification, together with contextual information such as temporality, certainty, and reason for the change (Mahajan et al., 2022, 2023).

Reliable extraction from unstructured clinical notes can support medication reconciliation and reduce medication-related harm (Killin et al., 2021), support pharmacovigilance and adverse drug event detection (Jagannatha et al., 2019; Henry et al., 2020), and provide structured inputs for clinical decision-support applications (Sutton et al., 2020).

However, medication information extraction from clinical text remains difficult because of abbreviations, shorthand, spelling variants, and domain-specific terminology. Determining medication status and temporality or linking attributes to the correct medication may also require surrounding context (Guo et al., 2024). Human-annotated reference corpora are therefore important for system development and rigorous evaluation.

Despite their importance, human-annotated clinical corpora remain scarce because privacy, ethical, and de-identification constraints restrict sharing beyond the institutions where the notes were produced (Wu et al., 2024). This limits reproducibility, direct system comparison, and assessment of generalisability across institutions.

MIMIC-IV-Note partially addresses the limited availability of clinical text by providing credentialed researchers with 331,794 de-identified dis-

charge summaries linked to the underlying MIMIC-IV electronic health record (Johnson et al., 2023). The core release, however, does not include human medication annotations. A separate resource provides medication labels for 600 discharge summaries, but they were generated using a large language model and their reported evaluation was conducted against the 2009 i2b2 corpus rather than against human annotations of the released MIMIC-IV notes (Goel et al., 2023b,a). Thus, no publicly released, human-annotated, medication-specific reference corpus is currently available for developing and evaluating medication information extraction systems on MIMIC-IV-Note.

To address this gap, we construct a human-annotated reference corpus of 305 MIMIC-IV-Note discharge summaries. The corpus captures medication mentions, their associated attributes and status, and the relations linking attributes to medications. Our contributions are:

- A fully adjudicated, human-annotated corpus of 305 MIMIC-IV-Note discharge summaries, supporting medication and attribute named-entity recognition, medication–attribute relation extraction, and medication-status classification across medication-list and narrative sections.
- A transparent annotation protocol and reproducible release, including a written annotation guideline, inter-annotator agreement analyses for entities, medication status, relations, and medication grouping, patient-disjoint data splits, and code for reproducing the corpus analyses.
- An empirical analysis of the distribution of medication mentions and attributes across discharge summaries, showing that medication-list sections and narrative sections contain comparable numbers of medication mentions but substantially different levels of attribute information.
- A reproducible supervised baseline for medication and attribute named-entity recognition that combines transformer encoders with a conditional random field, providing a reference point for future work.

These contributions provide, to our knowledge, the first human-annotated, medication-specific reference corpus for MIMIC-IV-Note, and a repro-

ducible benchmark for developing and evaluating medication information extraction systems on it.

2 Related Work

Medication information extraction has developed through a sequence of shared tasks and corpora that progressively expanded the representation of medication information in clinical text.

The 2009 i2b2 medication challenge established a major benchmark for medication information extraction from Partners HealthCare discharge summaries (Uzuner et al., 2010a). Its reference test set contains 251 summaries annotated for medication mentions, dose, route, frequency, duration, reason, and whether the information occurs in list or narrative text. The annotations were produced through a community process and achieved an F-measure above 0.90 against expert annotations (Uzuner et al., 2010b). The corpus established the core medication attributes adopted by many subsequent extraction tasks.

MADE 1.0 extended medication information extraction to pharmacovigilance using 1,089 longitudinal notes from 21 patients with cancer at the University of Massachusetts Memorial Hospital. The corpus annotates nine entity types and seven relation types, covering medications, indications, adverse drug events, and their associated attributes (Jagannatha et al., 2019).

The 2018 n2c2 adverse-drug-event and medication challenge introduced a corpus of 505 MIMIC-III discharge summaries annotated for nine concept types: drug, strength, form, dosage, frequency, route, duration, reason, and adverse drug event. It also defines eight relation types linking non-drug concepts to their associated medications (Henry et al., 2020). The corpus extends the i2b2 medication schema by integrating adverse-drug-event extraction.

The Contextualized Medication Event Dataset (CMED), used in the 2022 n2c2 shared task, extended medication information extraction to the contextual characterisation of medication changes (Mahajan et al., 2022, 2023). It uses notes from the 2014 i2b2/UTHealth corpus and annotates medication-change events by action, negation, temporality, certainty, and actor. CMED supports medication-timeline reconstruction but does not annotate medication attributes or link them to individual medication mentions.

Together, these corpora established the annota-

tion schemas and evaluation benchmarks that have supported medication information extraction. The i2b2, 2018 n2c2, and MADE corpora are based on source collections that predate the MIMIC-IV release. They therefore do not provide a human-annotated reference for evaluating medication information extraction on MIMIC-IV-Note.

MIMIC-IV-Note provides a more consistently maintained source of clinical text and remains available to credentialed researchers through PhysioNet (Johnson et al., 2023). For this collection, Goel et al. released i2b2-style medication labels for 600 discharge summaries, generated using a pipeline based on a large language model (Goel et al., 2023b,a). Their work demonstrates the potential of language models to reduce the human effort required for clinical information-extraction annotation. However, the reported medication evaluation uses the 2009 i2b2 reference, and no human evaluation of the released labels on MIMIC-IV-Note is reported.

To our knowledge, no publicly released human-annotated reference corpus has previously covered medication attributes and their relations in MIMIC-IV-Note discharge summaries. To address this gap, we develop a human-annotated corpus whose annotation schema builds on earlier medication information extraction corpora. It retains the core medication, dose, route, frequency, duration, and reason categories while extending the schema with allergy and medication status. Relative to i2b2, it analyses the document sections in which medications occur (medication-list versus narrative sections) rather than attaching a list-or-narrative field to each medication record, and uses explicit relations to link attributes to the medications they describe. Relative to the 2018 n2c2 schema, it adds an allergy entity and a medication-status attribute, and targets medication information rather than adverse drug events. Relative to CMED, it uses a three-valued medication-status attribute and adds the detailed medication attributes and medication-attribute relations that CMED does not annotate.

3 Corpus Construction

We constructed the corpus in five stages (Figure 1). We first selected documents by sampling from MIMIC-IV-Note, then developed the annotation schema and guideline. Annotation began with annotator training and proceeded to independent annotation. We then measured inter-annotator agreement on the independent annotations, and finally

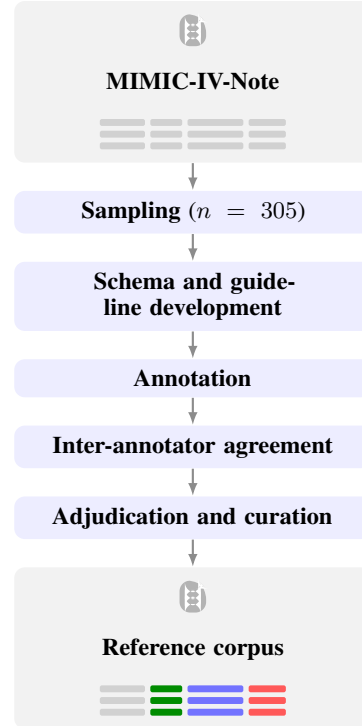


Figure 1: The corpus-construction pipeline.

adjudicated and curated them to produce the final corpus.

3.1 Data Source and Sampling

The corpus is drawn from MIMIC-IV-Note v2.2 (Johnson et al., 2023), which contains 331,794 de-identified discharge summaries from the Beth Israel Deaconess Medical Center in Boston.

We sampled from the full set of discharge summaries. Notes shorter than 500 characters were removed. We stratified the remaining notes on two axes, document length and medication density, each split into three bands at the tertiles of the eligible notes, to give a nine-cell grid, and drew a roughly equal number of notes from each cell so that short documents with many medications and long documents with few are not underrepresented. The discharge-medication section served as a proxy for medication density, since it could be counted before any annotation existed. The tertile cuts fell at 8,271 and 11,678 characters for length and at 7 and 13 entries for density. Notes with zero entries in the discharge-medication section were retained, since medications may appear elsewhere in the note. Sampling drew 33 notes per cell with a fixed seed of 42, the remaining 8 notes going to the largest cell, to form a final corpus of 305.

3.2 Annotation Schema and Guideline

Our schema builds on the 2009 i2b2 medication challenge and extends it.

Entity types. Seven entity types are annotated:

- **Medication:** the name of a drug, including brand names, generic names, and active ingredients.
- **Dose:** the amount of a medication given in a single administration.
- **Route:** how the medication is administered.
- **Frequency:** how often the medication is taken.
- **Duration:** how long the medication is taken.
- **Reason:** the clinical indication for the medication.
- **Allergy:** the drug or substance a documented allergy is to.

Medication status. Each medication carries a status attribute:

- **active:** the default, left blank;
- **discontinued:** the text states the medication was stopped or held;
- **neither:** the medication is named as past history rather than a current one, was considered but not prescribed or was explicitly rejected, or belongs to someone other than the patient.

Attribute relations. A medication’s dose, route, frequency, duration and reason are each linked to it by a relation: *has_dose*, *has_route*, *has_frequency*, *has_duration*, and *has_reason*. An attribute that applies to several medications is linked to each of them, so a single attribute span may carry more than one relation. An allergy is not a medication attribute, and therefore it carries no relation. Figure 2 shows a synthetic annotated example.

Guideline development. The annotation guideline was developed iteratively, beginning with the seven-field schema used in the 2009 i2b2 medication challenge. The lead annotator first annotated a small set of notes and revised the guideline to reflect the phenomena observed in the corpus. During annotator training, clarifying questions revealed cases that the guideline did not yet address. These

cases were incorporated into a final version that was fixed before the main annotation began.

The guideline defines each entity type through inclusion and exclusion criteria and specifies standard abbreviations for route and frequency, permitted status values, and relation types. Attributes are annotated only when they are explicitly stated in the text; annotators do not infer routes, frequencies, or other information that is not documented. Medication mentions are annotated in both medication-list and narrative sections. Each medication mention, together with its linked attributes, constitutes a medication group.

3.3 Annotation Workflow

Annotators. The corpus was annotated by four annotators with backgrounds in computer science. Because the task involved extracting information explicitly stated in the clinical notes, the annotation guideline assumed no prior medical or linguistic expertise. We adopted the schema used in the i2b2 medication challenge, which conducted a community annotation experiment and found that non-expert annotators achieved agreement comparable to that of experts. In that experiment, the community annotations agreed with expert annotations at F above 0.90 (Uzuner et al., 2010b).

Annotator training. Before independent annotation, the four annotators were trained on the guideline using a shared set of five notes. These notes were retained in the final corpus but excluded from the independent annotation and agreement analyses. The remaining 300 notes were then distributed among the annotators, of which 41 were independently annotated by more than one annotator (Section 3.4).

Annotation platform and recommender. Annotation was carried out in INCEpTION (Klie et al., 2018), a web-based annotation platform. For notes assigned to a single annotator, a string-matching recommender was enabled to accelerate annotation. The recommender proposed spans whose surface forms matched text that the same annotator had previously annotated. These suggestions were not applied automatically; each was accepted or rejected by the annotator.

The recommender was disabled for notes annotated independently by multiple annotators. These notes were used to measure inter-annotator agreement.

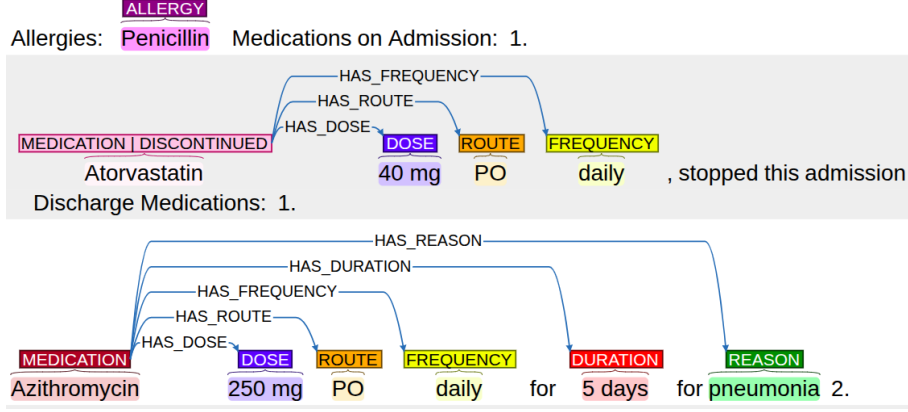


Figure 2: An example of an annotated note. The example is synthetic and contains no content from MIMIC.

3.4 Inter-Annotator Agreement

Agreement was measured from the independent annotations, before curation.

Metrics. Span agreement is reported as pairwise F-measure, since span-detection tasks provide no predefined set of negative units for estimating chance agreement (Hripcsak and Rothschild, 2005). For an entity type, the two annotators’ spans form sets A and B , with duplicates removed, and

$$F(A, B) = \frac{2 |A \cap B|}{|A| + |B|}. \quad (1)$$

Under exact agreement, spans match only at identical character offsets. For character-level agreement, each set is replaced by the character positions its spans cover,

$$\text{ch}(A) = \bigcup_{(b,e) \in A} \{b, b+1, \dots, e-1\}, \quad (2)$$

and F is computed over $\text{ch}(A)$ and $\text{ch}(B)$, giving partial credit when annotators mark the same mention but disagree on its boundaries.

F is computed per entity type for each annotator pair on each shared note. A type absent from both annotations is excluded; a type marked by only one scores zero. The overall figure pools counts across types within a comparison (an annotator pair on a shared note) before computing F , and comparisons are averaged with equal weight. Confidence intervals for exact agreement use 2,000 note-level bootstrap samples.

Status agreement is computed over medication spans both annotators marked at identical offsets; as these fix the units and the label set is closed, Cohen’s κ (Cohen, 1960) is reported with observed

agreement. Group agreement is F over (span, medication) pairs, each medication paired with itself and each attribute with its parent, so an attribute matches only when both annotators mark it identically and link it to the same medication. Attachment agreement is the proportion of attribute spans both marked and linked that they assign to the same medication, conditional on both having marked the span.

Results. Forty-one notes were annotated independently by more than one annotator: 15 by two, 10 by three, and 16 by all four. Each pair of annotators sharing a note yielded one comparison, for a total of 141 comparisons. Overall agreement was 0.896 under exact matching and 0.927 at the character level (Table 1). Exact agreement exceeded 0.87 for medication, dose, route, and frequency. Duration, reason, and allergy had lower estimates and wider confidence intervals. The character-level gains for duration and reason show that boundary differences account for some, but not all, of their disagreement.

Route was the only entity type for which character-level agreement was lower than exact agreement. This reflects the length weighting of the character-level measure: unmatched route spans were typically longer written-out forms, whereas matched spans were often short abbreviations.

Across the pairwise comparisons, 4,603 matched medication-span instances were available for status agreement: observed agreement was 0.977 and Cohen’s κ was 0.699. The lower κ reflects the strong class imbalance, with 95.9% of these instances labelled active.

Medication-group agreement was 0.872. Among attribute spans marked and linked by both annotators, 95.7% were assigned to the same medication. Thus, most group-level disagreement arose from

differences in span identification.

Entity type	Exact	95% CI	Char.	<i>n</i>
Medication	0.874	[0.844, 0.901]	0.916	41
Dose	0.946	[0.930, 0.961]	0.972	41
Route	0.925	[0.894, 0.952]	0.894	41
Frequency	0.892	[0.856, 0.925]	0.963	41
Duration	0.583	[0.441, 0.703]	0.657	21
Reason	0.566	[0.459, 0.674]	0.672	37
Allergy	0.653	[0.485, 0.839]	0.678	19
Overall	0.896	[0.876, 0.915]	0.927	41

Table 1: Pairwise inter-annotator agreement by entity type under exact and character-level matching; 95% bootstrap CIs.

3.5 Adjudication and Validation

Curation was performed using the INCEpTION curation view. For notes annotated by more than one annotator, annotations that matched exactly were treated as agreements, while discrepancies were highlighted for review. The lead annotator resolved these discrepancies according to the annotation guideline. Notes annotated by a single annotator were also reviewed against the guideline. This process produced one curated annotation layer for each note.

After curation, the corpus was checked using rule-based validation procedures derived from the guideline, and all flagged cases were reviewed manually and corrected where necessary. This produced the final corpus.

4 Corpus Analysis

4.1 Corpus Composition

The 305 notes contain 484,665 whitespace-delimited tokens, 35,097 entity spans, and 23,242 relations. Table 2 gives the distribution by entity type. Thus, although the corpus contains a relatively small number of documents, each note is annotation-dense, with a mean of 115.1 entity spans and 76.2 relations per note.

Medication mentions constitute 32.7% of all spans, with a mean of 37.6 per note; one note contains no medication mention. Duration and allergy together account for less than 2% of all spans. Of the 11,472 medication mentions, 370 are labelled discontinued and 290 neither; the remaining 94.2% have the default active status. Medication spans have a median of 1 token and dose spans 2, and no entity type has a median above 2.

Entity type	Spans	Per note
Medication	11,472	37.6
Dose	8,740	28.7
Frequency	6,764	22.2
Route	6,009	19.7
Reason	1,416	4.6
Allergy	420	1.4
Duration	276	0.9
All spans	35,097	115.1
Relations	23,242	76.2

Table 2: Corpus composition.

4.2 Medication Information by Document Section

Document sections were identified using corpus-derived header patterns. Candidate headers were strings ending in a colon at the beginning of a line or following two spaces. We distinguished section headers from internal subheadings by the presence of a preceding blank line: candidates were preceded by a blank line in over 62% of their occurrences or in under 36%, with none in between, so this cutoff was set by the corpus itself. This procedure yielded 25 recurring section headers, including explicit variants for headers partially affected by de-identification. Variant headers were mapped to canonical section labels, and every medication mention was assigned to one of the resulting sections; none occurred before the first recognised header.

Of all medication mentions, 55.4% occur in the two medication-list sections, medications on admission and discharge medications, and 44.6% in the narrative sections. The largest concentrations outside the lists are in the brief hospital course (18.8%), history of present illness (7.8%), pertinent results (6.9%), discharge instructions (6.8%), and past medical history (3.8%).

The number and type of attributes linked to a medication differ substantially between medication-list and narrative sections. Overall, 31.3% of medication mentions have no linked attributes, and this proportion differs between the two (Table 3).

In the medication-list sections, a medication mention carries a mean of 2.86 distinct attribute types, usually dose, route, and frequency; in the narrative sections the mean is 0.59, and 65.5% of mentions are bare medication names. Detailed medication information is thus concentrated in the medication-list sections, even though the two contain comparable numbers of mentions.

Attribute types on a medication	List	Narrative
0 (bare medication name)	3.8%	65.5%
1	3.7%	17.2%
2	10.7%	11.2%
3	66.9%	5.1%
4 or 5	14.9%	0.9%
Mean	2.86	0.59
Medications	6,361	5,111

Table 3: Attribute types per medication mention in list and narrative sections.

5 Baseline Experiments

5.1 Data

We evaluated on a fixed training, development, and test partition, summarised in Table 4, which we plan to release with the corpus. The partition is stratified using the length and medication-density grid used to sample the corpus, preserving a comparable mix of document types across splits. It is also patient-disjoint: all notes from a patient are assigned to the same split, preventing evaluation on patients represented in the training data. The five shared annotator-training notes are included in the training set.

Split	Notes	Tokens	Medications	Spans
Train	185	292,000	6,794	20,860
Dev	62	99,163	2,534	7,526
Test	58	93,502	2,144	6,711

Table 4: Fixed splits, stratified and patient-disjoint.

5.2 Setup

We established a supervised baseline for medication and attribute named-entity recognition using a transformer encoder with a conditional random field over BIO tags. Notes exceeding the encoder limit are divided into 512-token windows that overlap by 128 tokens. For tokens appearing in multiple windows, predictions are retained from the window in which the token is furthest from a boundary, and character-level spans are reconstructed once per note.

We fine-tuned Bio_ClinicalBERT (Alsentzer et al., 2019) and PubMedBERT (Gu et al., 2021) using AdamW at a learning rate of 3×10^{-5} and an effective batch size of 32 (batch size 8 with gradient accumulation of 4), for up to 15 epochs; checkpoint selection and early stopping (patience 3) used the development-set exact-span F-measure. Performance was evaluated on the held-out test set using

the exact and character-level criteria defined in Section 3.4, with counts pooled across notes. Each configuration was run with three random seeds; we report the mean across seeds, with the standard deviation for the overall figures.

5.3 Results

The two encoders were comparable, with PubMedBERT marginally higher overall (Table 5): mean exact-span and character-level F-measures of 0.945 and 0.967, compared with 0.941 and 0.966 for Bio_ClinicalBERT. Exact-span F-measure exceeded 0.93 for medication, dose, route, and frequency with both encoders. Performance was lower for duration and reason, and PubMedBERT was higher on reason and allergy.

The baseline exact-span F-measure was higher than the pairwise inter-annotator agreement reported in Section 3.4, but the values are not directly comparable. The model was evaluated against a single adjudicated reference with counts pooled across the test set, whereas agreement was averaged over pairwise note-level comparisons between independent pre-adjudication annotations.

5.4 Data-Split Sensitivity and Learning-Curve Analysis

To assess sensitivity to the fixed partition, we re-generated the patient-disjoint, stratified 60/20/20 split with ten random seeds, including the released split, and retrained both encoders on each. The training seed was held fixed, so variation reflects split choice rather than the training-seed variation reported in Table 5. Across splits (Figure 3), mean exact-span F-measure was 0.936 (95% CI [0.931, 0.941]) for Bio_ClinicalBERT and 0.940 ([0.938, 0.943]) for PubMedBERT, with ranges of 0.922–0.942 and 0.931–0.944, respectively. PubMedBERT was marginally but significantly higher (mean difference 0.004, 95% CI [0.001, 0.007]; two-sided paired t -test, $t(9)$, $p = 0.010$).

We also examined whether the corpus is large enough to train a model by retraining on 25, 50, 75, and 100% of the training set with development and test fixed (Figure 4). Performance improved most at smaller training sizes and then flattened. From 139 to 185 notes, exact-span F-measure increased from 0.936 to 0.941 for Bio_ClinicalBERT and from 0.938 to 0.944 for PubMedBERT. Although the curve has not fully saturated, the corpus is already in a region of diminishing returns.

Entity type	Bio_ClinicalBERT		PubMedBERT	
	Exact	Char.	Exact	Char.
Medication	0.938	0.975	0.939	0.976
Dose	0.949	0.978	0.952	0.981
Route	0.970	0.961	0.972	0.962
Frequency	0.963	0.980	0.964	0.979
Duration	0.719	0.791	0.701	0.707
Reason	0.756	0.829	0.796	0.838
Allergy	0.916	0.956	0.951	0.972
Overall	0.941 ± 0.001	0.966 ± 0.001	0.945 ± 0.003	0.967 ± 0.001

Table 5: Baseline entity-recognition F-measure on the 58-note test set, averaged over three seeds.

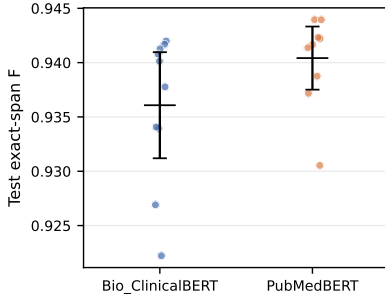


Figure 3: Exact-span F-measure across ten patient-disjoint stratified splits; bars show 95% CIs.

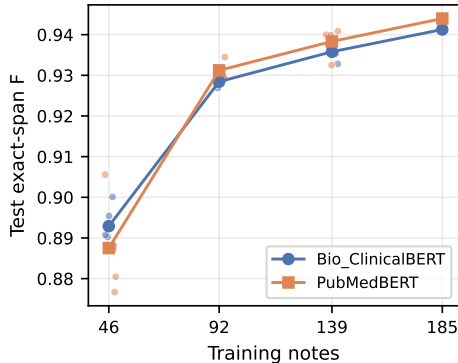


Figure 4: Exact-span F-measure by training-set size, with development and test sets fixed.

5.5 Error Analysis

We compared predicted and reference spans on the test set, categorising non-exact matches as boundary, type, missed, or spurious errors (Table 6). Type errors were rare (7 for each encoder), while most errors involved span detection or boundary selection. Boundary errors accounted for 168 of 306 non-exact reference spans for Bio_ClinicalBERT and 173 of 364 for PubMedBERT, which is consistent with the higher character-level than exact-span agreement observed between annotators.

The encoders had similar overall F-measures but different error patterns. Bio_ClinicalBERT produced more spurious and fewer missed spans, whereas PubMedBERT showed the reverse pattern. Errors were relatively uncommon for medication, dose, route, and frequency, but reason and duration remained more difficult. Thus, the high aggregate F-measure is driven mainly by frequent entity types, while performance on the rarer reason and duration categories remains less reliable, indicating that extraction of these more context-dependent attributes is not yet solved.

6 Discussion

The corpus demonstrates that medication information in discharge summaries is not confined to dedicated medication-list sections. The two medication-list sections and the narrative sections contain comparable numbers of medication mentions, but the information associated with those mentions differs substantially. Mentions in the lists contain a mean of 2.86 attribute types, compared with 0.59 in the narrative. A system restricted to the medication-list sections would therefore recover only about half of all medication mentions. Although it would capture most mentions documented with a complete dose, route, and frequency, it would miss those recorded elsewhere, particularly in the brief hospital course, history of present illness, and discharge instructions sections. Because medication reconciliation and related safety tasks require a complete medication record, the lists alone are insufficient. By annotating medication mentions wherever they occur, the corpus supports the development and evaluation of systems that process the entire note.

The agreement and baseline results together clarify which aspects of the task are well defined and which remain challenging. Agreement is high for medication, dose, route, and frequency, which are

Entity type	Ref.	Bio_ClinicalBERT				PubMedBERT			
		Bound.	Type	Miss	Spur.	Bound.	Type	Miss	Spur.
Medication	2,144	54	1	48	90	60	0	57	93
Dose	1,685	52	1	6	62	55	1	18	54
Route	1,163	11	1	19	32	7	1	30	19
Frequency	1,309	20	3	13	51	16	4	23	46
Duration	56	4	0	7	16	2	1	17	7
Reason	295	26	1	38	46	32	0	39	35
Allergy	59	1	0	0	10	1	0	0	3
Overall	6,711	168	7	131	307	173	7	184	257

Table 6: Test-set error counts by entity type. Ref.: reference; Bound.: boundary; Type: type; Miss: missed; Spur.: spurious.

frequent and have short, clearly bounded surface forms. It is lower for duration, reason, and allergy, which occur less often and have less clearly defined span boundaries. The improvements observed under character-level evaluation for duration and reason suggest that some disagreement concerns span extent rather than the underlying entity identity. The supervised baseline follows the same pattern: it achieves exact-span F-measures above 0.93 for the frequent entity types but performs substantially worse on duration and reason. These categories therefore appear to be intrinsically difficult for both annotation and modelling.

Our baseline covers entity recognition; the relation- and status-classification tasks the corpus defines, together with a direct evaluation of large language model annotations against the human reference, remain for future work.

A cross-dataset evaluation against an established medication corpus such as the 2018 n2c2 dataset would provide additional evidence about how well models trained on this corpus generalise across clinical-note collections and remains an important direction for future work.

7 Conclusion

We introduced the first human-annotated corpus for medication information extraction from MIMIC-IV discharge summaries. We reported its inter-annotator agreement and a supervised baseline that achieved an exact-span F-measure of 0.945 on the held-out test set, and we plan to release the corpus through PhysioNet together with its annotation guideline, patient-disjoint data splits, and analysis code. Unlike resources restricted to dedicated medication-list sections, it annotates medication mentions and their attributes throughout the entire note. By providing an in-domain human reference

where previous resources relied on automatically generated labels, the corpus will enable the development, evaluation, and validation of medication information extraction systems for MIMIC-IV-Note. In turn, these systems can support medication reconciliation and other safety-critical applications aimed at reducing medication errors and preventable patient harm.

Limitations

The corpus is derived from a single hospital system, is limited to English-language discharge summaries, and may therefore not generalise to other institutions, note types, or languages. In addition, the sample was stratified rather than designed to be representative of MIMIC-IV-Note as a whole. Consequently, the observed entity frequencies and the proportion of medication mentions occurring in medication lists (55.4%) should be interpreted as properties of this sample rather than population estimates.

We intend the corpus, together with its annotation guideline, data splits, and analysis code, to serve as a reproducible basis for extension to further institutions, note types, and languages.

Access and Licensing

We plan to release the corpus through PhysioNet under the PhysioNet Credentialed Health Data License, following the access conditions for MIMIC-IV-Note.

Ethics Statement

The corpus is a secondary annotation layer over MIMIC-IV-Note, a de-identified clinical database whose collection and release received institutional review board approval and whose distribution is

managed by PhysioNet. Access requires PhysioNet credentialing and acceptance of the Credentialed Health Data Use Agreement; any release of the annotations would inherit these conditions and go only through the same repository. The annotation work used the de-identified text as distributed, made no attempt to re-identify patients, and reproduces no note text in this paper, where any examples used are synthetic.

Acknowledgments

The authors would like to thank the data annotators for their time, and valuable contributions to the annotation process. The authors also acknowledge the anonymous reviewers for their valuable comments and suggestions, which helped improve the quality of this manuscript.

This research was funded through the CREATE-DkIT project, supported by the HEA TU-Rise program and co-financed by the Government of Ireland and the European Union through the Southern, Eastern & Midland Regional Program of the ERDF 2021-27 and the Northern & Western Regional Programme 2021-27.



References

- Emily Alsentzer, John Murphy, William Boag, Weihung Weng, Di Jindi, 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.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- Akshay Goel, Almog Gueta, Omry Gilon, Sofia Erell, and Amir Feder. 2023a. [Medication Extraction Labels for MIMIC-IV-Note Clinical Database](#). *PhysioNet*. Version 1.0.0.
- Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. 2023b. [Llms accelerate annotation for medical information extraction](#). In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 82–100. PMLR.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. [Domain-specific language model pretraining for biomedical natural language processing](#). *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.
- Yuting Guo, Yao Ge, and Abeed Sarker. 2024. [Detection of medication mentions and medication change events in clinical notes using transformer-based models](#). In *Studies in health technology and informatics*, volume 310, page 685.
- 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.
- George Hripcsak and Adam S Rothschild. 2005. [Agreement, the f-measure, and reliability in information retrieval](#). *Journal of the American medical informatics association*, 12(3):296–298.
- Abhyuday Jagannatha, Feifan Liu, Weisong Liu, and Hong Yu. 2019. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (made 1.0). *Drug safety*, 42(1):99–111.
- Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. [MIMIC-IV-Note: Deidentified free-text clinical notes](#). *PhysioNet*. Version 2.2.
- Lauren Killin, Areej Hezam, Kelly K Anderson, and Blayne Welk. 2021. [Advanced medication reconciliation: a systematic review of the impact on medication errors and adverse drug events associated with transitions of care](#). *The Joint Commission Journal on Quality and Patient Safety*, 47(7):438–451.
- Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. [The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation](#). In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.
- Diwakar Mahajan, Jennifer J Liang, and Ching-Huei Tsou. 2022. Toward understanding clinical context of medication change events in clinical narratives. In *AMIA Annual Symposium Proceedings*, volume 2021, page 833.
- Diwakar Mahajan, Jennifer J Liang, Ching-Huei Tsou, and Özlem Uzuner. 2023. [Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes](#). *Journal of biomedical informatics*, 144:104432.
- Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. [An overview of clinical decision support systems: benefits, risks, and strategies for success](#). *NPJ digital medicine*, 3(1):17.

- Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010a. [Extracting medication information from clinical text](#). *Journal of the American Medical Informatics Association*, 17(5):514–518.
- Özlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. 2010b. [Community annotation experiment for ground truth generation for the i2b2 medication challenge](#). *Journal of the American Medical Informatics Association*, 17(5):519–523.
- Jiageng Wu, Xiaocong Liu, Minghui Li, Wanxin Li, Zichang Su, Shixu Lin, Lucas Garay, Zhiyun Zhang, Yujie Zhang, Qingcheng Zeng, Jie Shen, Changzheng Yuan, and Jie Yang. 2024. [Clinical text datasets for medical artificial intelligence and large language models: A systematic review](#). *NEJM AI*, 1(6).