Med2Story Referential: A Domain-Specific Extension of ISO 24617-9 for Clinical Narratives Annotation

Ana Luísa Fernandes[*‡‖], **Purificação Silvano**[*‡‖]
Nuno Guimarães[*†], **Luís Filipe Cunha**[*†] **Rita Rb-Silva**[§¶], **Alípio Jorge**[*†]
[*] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC)
R. Dr. Roberto Frias, Porto, Portugal
{ana.l.fernandes, purificacao.silvano, nuno.r.guimaraes, luis.f.cunha, alipio.jorge}@inesctec.pt
[†] Faculty of Sciences, University of Porto
Rua do Campo Alegre, Porto, Portugal
[‡]Faculty of Arts and Humanities, University of Porto
Via Panorâmica, Porto, Portugal
[‖]Centre for Linguistics of the University of Porto (CLUP)
Via Panorâmica, Porto, Portugal
[§] CI-IPOP
Rua Dr. António Bernardino de Almeida, Porto, Portugal
[¶] MedTechnologist
Rua Dr. Manuel de Arriaga, Valongo, Portugal
rrsilva@med.up.pt

Abstract

The semantic annotation of clinical narratives is particularly challenging due to the complexity of medical discourse and the need to integrate linguistic, semantic, and domain-specific information within a unified framework. Existing schemes tend to fall into two categories: general-purpose frameworks, which offer robust linguistic modelling but lack specialised medical representation, and domain-specific schemes, which capture clinical content yet often fail to distinguish fundamental semantic types, especially eventive expressions and referential entities. To address this gap, this study proposes Med2Story Referential, a new extension of the Text2Story annotation scheme (Silvano et al., 2021; Leal et al., 2022) (based on ISO 24617-9: 2019) dedicated to referential entities in clinical narratives. Building on previous work that introduced a specialised branch for eventive entities (Fernandes et al., 2025a), and informed by the UMLS Metathesaurus and expert validation from a consultant haematologist, the extension introduces eight referential categories that refine the representation of clinical actors, substances, biological entities, instruments, and documentation. The results show that ISO 24617-9: 2019 can be applied to this type of text; however, several adaptations are required, particularly with regard to the grammatical domain and the inclusion of specialised domain labels. Nonetheless, the annotation experiment conducted to validate our proposal showed that the annotation scheme and its accompanying guidelines enable a comprehensive and detailed representation of both grammatical and medical aspects. Moreover, the results indicate that the scheme can be applied effectively by annotators without medical expertise.

Keywords: clinical narratives, Electronic Health Records, annotation scheme, referential annotation, health data

1. Introduction

Electronic Health Records (EHRs) constitute a rich source of clinical information and may play a decisive role in addressing issues related to healthcare quality, clinical decision support, and the reliable flow of information among individuals and departments involved in patient care (Kundeti et al., 2016). In this context, Natural Language Processing (NLP) techniques play a fundamental role by transforming unstructured data into structured data, thereby enabling systematic exploration. Information extraction may be carried out through methods such as Named Entity Recognition (NER) and Relation Extraction (RE), which are central NLP components applied to the clinical domain (Durango et al., 2023). These tasks allow for the identification, extraction, and classification of clinical terms, such as diseases, symptoms, treatments, tests, pharmaceuticals, procedures, and body parts, thereby enabling the automatic recognition of a wide range of clinical concepts (Durango et al., 2023). However, these concepts differ substantially from both a linguistic and ontological perspective, which poses significant challenges to their annotation and computational modelling. For instance, diseases and symptoms correspond semantically to states; treatments, tests, and procedures correspond to events; and pharmaceuticals and body parts to entities with referential existence. Capturing these distinctions in a consistent and operational way requires annotation frameworks capable of representing the

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 1–12
12 May 2026.

semantic and referential properties of clinical entities in a structured manner.

Manual annotation constitutes one of the fundamental pillars of linguistic research and NLP (cf. Snow et al., 2008, Flickinger et al., 2017). In addition to enabling the empirical study of linguistic phenomena, it provides essential reference labels for the training and evaluation of language models across diverse tasks (cf.Pustejovsky and Stubbs, 2012, Pustejovsky et al., 2017, Levi and Shenhav, 2022). It also contributes to the formalisation of linguistic theories by offering a structured framework for their empirical validation (Hovy and Lavid, 2010). The development of robust and coherent annotation schemes is therefore of particular relevance: a well-designed scheme should ensure systematicity, consistency, interoperability, and comprehensiveness, thereby allowing complex linguistic phenomena to be represented in a theoretically grounded and operationally applicable manner by annotators (Beck et al., 2020).

In the clinical domain, existing annotation schemes tend to treat uniformly entities that differ significantly in their linguistic and ontological nature (cf. Campillos-Llanos et al., 2018, Sun et al., 2013, Kundeti et al., 2016). For instance, the i2b2 temporal annotation scheme classifies drugs (e.g., *Diltiazem*) as events, even though they correspond to participants rather than events in linguistic terms. This uniform treatment of ontologically heterogeneous entities hinders the rigorous establishment of referential and anaphoric relations, as well as the construction of a coherent temporal line of clinical events and states. It is therefore essential to adopt consistent semantic principles and linguistic conventions that facilitate the automatic extraction of information and its holistic interpretation since studies have shown significant improvements in clinical information processing resulting from the use of appropriate annotation frameworks (Deléger et al., 2017).

Among existing schemes, Text2Story (Silvano et al., 2021, Leal et al., 2022) stands out as a generalist annotation scheme based on ISO 24617 and adapted to European Portuguese, proposing a layered annotation approach encompassing temporal, referential, spatial, and semantic role layers. The scheme has been successfully applied to the annotation of journalistic narratives in Portuguese (Silvano et al., 2023), literary narratives in Portuguese, English, and Spanish (Cantante et al., 2024), and financial narratives in Portuguese (Leal et al., 2025). Moreover, the temporal layer of the scheme has already been adapted to clinical narratives, resulting in the development of Med2Story Temporal (Fernandes, 2025). In Text2Story, events, i.e., eventive entities that denote something that occurs (e.g., “the patient underwent a full blood count.), are annotated in the temporal layer, whereas participants, i.e., entities with referential status that play an important role in the narrative (e.g., “the full blood count was added to the patient’s electronic record”), are annotated in the referential layer. Each layer includes different labels and attributes that enable a general semantic and morphosyntactic characterisation of the represented entities and the relations between them. While these tags may be suitable for representing general narrative structures, clinical narratives pose additional challenges due to the density and specificity of medical information they contain. In particular, medical texts frequently refer to clinical procedures, examinations, conditions, and treatments whose interpretation requires a more fine-grained representation of their semantic and grammatical properties. The current referential layer of Text2Story does not provide sufficiently detailed labels and attributes to capture this type of domain-specific information in a consistent and structured way. As a consequence, the representation of medical referential entities remains limited, which may affect downstream tasks such as clinical information extraction, normalization, and interoperability with medical knowledge resources.

In this work, we address this limitation by investigating how the referential representation can be extended to better accommodate the specific characteristics of clinical narratives. More specifically, we propose to: (i) assess the extent to which the ISO 24617-9: 2019 standard is suitable for representing the referential structure of clinical narratives in European Portuguese; and (ii) define and integrate an extension of the referential layer of the Text2Story annotation scheme (based on ISO 24617-9: 2019) that enables the consistent representation of medical and grammatical information (Med2Story Referential).

Our main contributions are the following: (i) a referential annotation framework for clinical data based on ISO 24617-9: 2019; (ii) the validation of the proposed annotation framework through an annotation experiment.

The remainder of the article is structured as follows. Section 2 reviews the related work. Section 3 presents the framework used for referential annotation: in Section 3.1, we describe the methodology adopted to develop the extension of the annotation scheme, and in Section 3.2, we introduce the proposed Med2Story Referential framework. Section 4 presents the experimental study: Section 4.1 outlines the research design, while Section 4.2 reports the results of the annotation experiment. Finally, Section 5 presents the conclusions and future work, and Section 6 discusses the limitations of the study.

2. Related work

In the clinical domain, annotation guidelines assume particular importance for the preparation of structured data and for the representation of textual content for human interpretation and information retrieval. They also support the training of supervised learning models and enable the comparative evaluation of language technologies through reference datasets (gold standards). Clinical annotation tasks are, however, especially demanding due to the heterogeneity of clinical texts and the breadth of knowledge they encompass, spanning multiple dimensions such as diseases, signs, symptoms, clinical findings, and procedures, as well as contextual factors including temporality, factuality, and other relevant interpretive parameters (Schulz et al., 2023). Despite this complexity, many existing annotation guidelines have been developed primarily within the context of shared tasks aimed at addressing specific NLP, such as NER and RE (cf. Harnoune et al. (2021); Yao et al. (2015); Miranda-Escalada et al. (2020)). This orientation has often resulted in relatively shallow annotation policies focused on immediate application-driven objectives rather than on comprehensive linguistic and semantic modelling (Schulz et al., 2023).

Several annotation frameworks adopt a layered architecture that formally distinguishes between entities such as participants and events. Examples include the ISO 24617 standards (ISO 24617-1: 2012; ISO 24617-9: 2019) and the Text2Story annotation scheme (Silvano et al., 2021; Leal et al., 2022). While such general-purpose frameworks enable fine-grained morphosyntactic and semantic analysis, they remain largely domain-independent and may therefore fail to capture specialised information required in highly technical fields such as medicine.

With regard to domain-specific annotation, a number of frameworks have been developed, particularly for the temporal dimension of clinical texts. Notable examples include THYME-TimeML (Styler IV et al., 2014), an adaptation of TimeML (Pustejovsky et al., 2003) tailored to clinical narratives. Other TimeML-inspired initiatives include the i2b2 project for English clinical records (Sun et al., 2013), the MERLOT corpus for French (Campillos-Llanos et al., 2018), and MEDSPANER for Spanish (Campillos-Llanos et al., 2025). In addition, recent shared tasks, such as the 2024 Chemotherapy Treatment Timeline Extraction challenge (Yao et al., 2024), aim to advance the automatic extraction of clinical event timelines from EHRs. González-Moreno et al. (2025) proposed an annotation scheme for Spanish clinical narratives concerning nut allergy; however, they do not distinguish between eventive entities (e.g. comorbidities such as "asthma") and purely referential entities (e.g. allergies such as "cat or dog"). More broadly, existing schemes frequently lack clearly differentiated annotation layers and tend to conflate distinct semantic categories, including eventive expressions and referential entities. This limitation reduces analytical precision and constrains interoperability across annotated resources.

General-purpose medical annotation tools have increasingly incorporated functionalities for the recognition and classification of medical entities, as exemplified by MedLEE (Friedman, 2000), MetaMap (Aronson and Lang, 2010), and MetaMap Lite (Demner-Fushman et al., 2017), as well as mechanisms for negation and assertion detection, such as those implemented in cTAKES (Savova et al., 2010) and CLAMP (Soysal et al., 2018). With respect to NER strategies, these tools frequently rely on dictionary-based approaches, which enable the identification of entities in unsupervised contexts and effectively handle infrequent mentions (cf. Campillos-Llanos et al., 2025). Such approaches have been combined with rule-based and machine learning methods and, more recently, with neural models adapted or pre-trained on large volumes of biomedical and clinical data (Campillos-Llanos et al., 2025). However, they remain insufficient for capturing specialised information concerning discourse-relevant referential entities.

To the best of our knowledge, no existing annotation scheme simultaneously satisfies three key requirements: (i) an explicit semantic distinction between eventualities (events and states) and participants (discourse-relevant referring expressions); (ii) a rigorously layered architecture integrating multiple linguistic levels; and (iii) the systematic representation of grammatical, semantic, and domain-specific medical information within a unified framework. This gap highlights the need for a comprehensive annotation model capable of capturing the full complexity of clinical narratives while remaining interoperable, linguistically grounded, and computationally exploitable.

3. Framework for Referential Annotation

3.1. Methodology for Building the Referential Layer

This study aims to assess the applicability of the ISO 24617-9: 2019 framework to the referential annotation of clinical narratives in Portuguese and to develop an extension of the referential component of the Text2Story annotation scheme specifically designed for clinical narratives, hereafter *Med2Story Referential*. To this end, we adopted the four-phases methodological framework proposed by Fer-

nandes et al. (2025b), originally used in the development of the temporal-layer extension of Text2Story for clinical narratives (*Med2Story Temporal*).

Phase 1 – Foundational Preparation: We began with a literature review of annotation schemes capable of supporting referential annotation in clinical narratives. This review showed that most existing schemes do not systematically distinguish between eventive nouns and nouns with referential existence (cf. MERLOT (Campillos-Llanos et al., 2018) and i2b2 (Sun et al., 2013)), and, as previously observed by Fernandes (2025), clinical-domain labels tend to be overly broad. In light of these findings, the referential layer of the Text2Story annotation scheme was selected as the base framework due to its comprehensiveness, interoperability, language independence, and ability to integrate annotations across multiple semantic layers within a harmonised architecture. Furthermore, Text2Story derives from an adaptation of the ISO 24617 framework, incorporating adjustments tailored to European Portuguese, making it particularly suitable for the linguistic characteristics of the dataset analysed. Its interoperability has already been demonstrated through successful adaptations to multiple domains, providing further evidence of its flexibility and robustness while preserving a coherent annotation structure and supporting domain-specific extensions and specialised analytical requirements. Having selected the base framework, we conducted an experiment to identify strengths and weaknesses of the scheme. This involved manual annotation of five pseudonymised medical reports derived from group consultations with five patients diagnosed with Acute Myeloid Leukemia (AML) and followed at IPO-Porto, Portugal. These reports were annotated according to the referential layer of Text2Story. The annotation focused on participants, i.e., discourse-relevant referential expressions.

Phase 2 – Design and Specification: The resulting dataset was analysed to characterise referential expressions and to identify the adaptations required for the Text2Story scheme to adequately represent clinical narratives. To ensure sufficient domain coverage and determine the full set of required medical-domain labels, a broader exploratory analysis was conducted on a larger corpus consisting of 38 pseudonymised real medical reports from nine AML patients followed at the same institution. This extended analysis enabled the identification of additional domain-specific categories and ensured that the proposed extension captured all relevant referential phenomena present in clinical discourse. The proposed extension is presented in 3.2.

Phase 3 – Empirical Validation: A pilot annotation study was conducted with two annotators and one curator. Annotation results were analysed with particular attention to inter-annotator agreement (IAA), sources of disagreement between annotators and the curator, and qualitative feedback. Annotators also evaluated the scheme and its guidelines using Likert-scale questionnaires. Based on these results, refinements were introduced to both the annotation scheme and the guidelines. Findings from this phase are reported in 4.

Phase 4 – Consolidation and Refinement: Finally, the annotation scheme and its guidelines were consolidated and further improved in light of the empirical results obtained in Phase 3.

3.2. Med2Story Referential Proposal

The referential layer of Text2Story, grounded in ISO 24617-9: 2019, focuses on the identification of referring expressions in the text, the determination of the discourse entities they denote, and the specification of relations among those entities (Silvano et al., 2021). With regard to Text2Story and specifically to discourse-entity structure, the attribute *Lexical_Head* was considered, originally comprising the values *Noun* and *Pronoun*. For referential-expression structure, the attributes *Domain* and *Involvement* were incorporated. The former encodes *Individuation*, with the values *Set*, *Individual*, or *Mass*, and includes the attribute *Type*, whose values derive from named-entity classification typologies, namely *PER* (person), *ORG* (organisation), *LOC* (location), *OBJ* (object), *NAT* (nature), and *OTHER*. The *Involvement* attribute specifies the degree of participant engagement in an event, with the values 0, 1, >1, all, or undefined, as defined in the scheme (Silvano et al., 2021).

To represent nominal anaphora mechanisms, Text2Story also employs the objectal links defined in ISO 24617-9, enabling the identification of relations such as objectal identity, part–whole (part_of), membership (member_of), and referential disjunction (Silvano et al., 2021).

Following the experimental annotation carried out in Phase 1, it became evident that certain modifications to the Text2Story referential layer were required in order to enhance its informativeness for capturing both grammatical and domain-specific medical information in clinical narratives. Concerning the *Lexical_Head* attribute, two additional values were introduced: *adverb*, to capture instances such as *Foi transferida para o IPO, onde foi confirmado o diagnóstico.* [“She was then transferred to the IPO, where the diagnosis was confirmed.”], and *adjective*, to account for expressions such as *infeção urinária* ["urinary infection]. Although the head of *infeção urinária* is the noun *infeção*, analysing the adjective as part of the event markable would result in the loss of essential medical information, namely the location of the infection. A similar approach was adopted by Campillos-Llanos et al. (2018), who annotate adjectives such as “bilat-

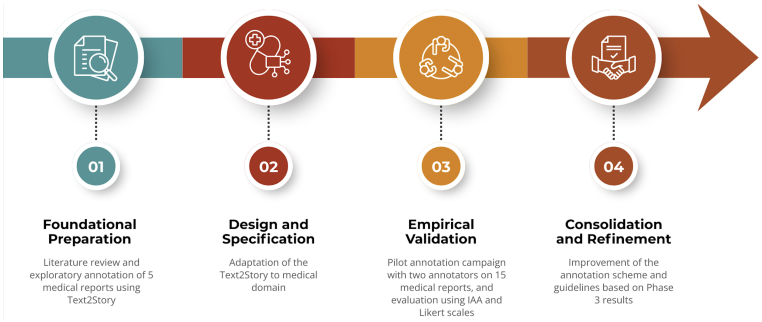

Figure 1: Methodology used for the development and validation of the annotation scheme, based on Fernandes et al. (2025b).

eral" separately from the noun phrase to which they belong. With regard to the *Involvement attribute*, the original set of values defined in the scheme was retained. In contrast, most of the values originally provided for the *Type* attribute proved unsuitable for the medical domain, resulting in many participants remaining unclassified. This observation indicated the need for a comprehensive revision and domain-specific adaptation of the attribute set. To this end, a larger corpus comprising 38 medical reports of different types was re-examined. In order to determine the medical categories required to capture the relevant information contained in the reports, new attributes were proposed inductively as participants emerged in the corpus, using the UMLS Metathesaurus (Bodenreider, 2004) as a conceptual reference framework.

The resulting inventory was organised into eight major categories, each comprising top-level attributes (Figure 2) and second and third-level attributes (Figure 3).

Because numerous nominal expressions with eventive semantics may also receive referential interpretations depending on context, the study additionally involved the development of a set of diagnostic tests designed to distinguish eventive from referential entities. These tests were developed based on the works of Grimshaw (1990), Brito and Oliveira (1997), Borer (2003), and Roy and Soare (2013). They were incorporated into the annotation guidelines to support annotators in consistently differentiating between the two categories and thereby improve annotation reliability.

Regarding the OLINKs, we noticed that they also required significant modifications, which will be addressed in future work. In this study, the focus will be on the participants and their attributes.

In Figure 4, we present an example of a medical report annotated according to the *Med2Story Referential* annotation scheme.

The development of Med2Story Referential was undertaken by a linguist with a background in Pharmaceutical Sciences. As argued by Roberts et al. (2009), the integration of linguistic expertise with

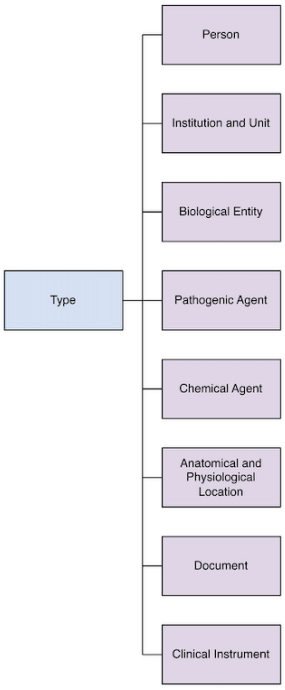

Figure 2: Top-level attributes of *Type*

clinical knowledge in the annotation process tends to enhance annotation quality and reliability. In line with this perspective, the linguistic labels were validated by two specialists from distinct areas of Semantics, while the medical-domain labels were validated by a physician specialising in Haematology. The Med2Story Referential guidelines used for the pilot-study annotation of Phase 3, can be consulted in the the project's GitHub Repository[1].

[1] https://github.com/analuisacardosofernandes/Med2Story-Referential/

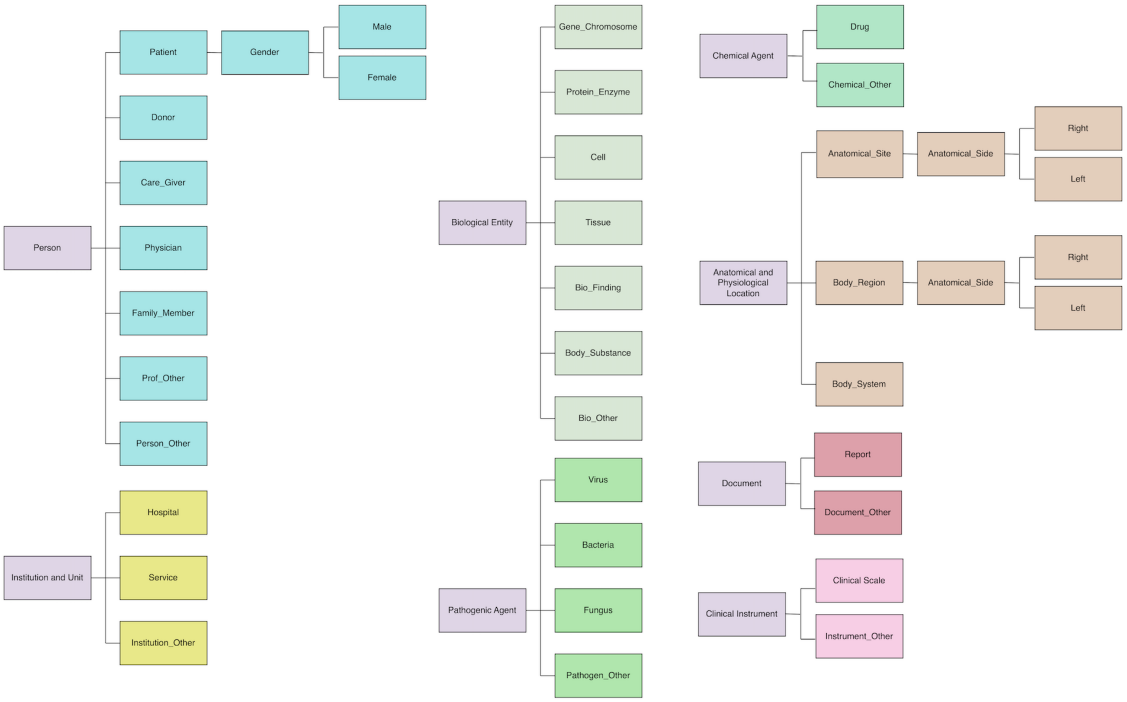

Figure 3: Second and third-level attributes of *Type*

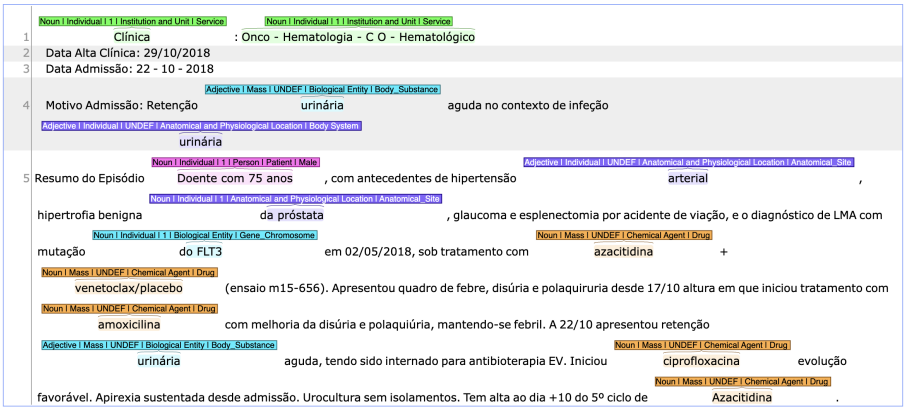

Figure 4: Excerpt from a medical report annotated according to the *Med2Story Referential* annotation scheme. *Episode Summary: A 75-year-old patient with a medical history of arterial hypertension, benign prostatic hyperplasia, glaucoma, and splenectomy following a road traffic accident, and a diagnosis of AML with an FLT3 mutation on 02/05/2018, undergoing treatment with azacitidine + venetoclax/placebo (trial M15-656). The patient presented with fever, dysuria, and pollakiuria since 17/10, when treatment with amoxicillin was initiated, resulting in improvement of dysuria and pollakiuria, although the fever persisted. On 22/10 the patient developed acute urinary retention and was admitted for intravenous antibiotic therapy. Ciprofloxacin was initiated, with favourable clinical evolution. Sustained apyrexia since admission. Urine culture showed no bacterial growth. The patient was discharged on day +10 of the 5th cycle of azacitidine.*

4. Experimental Study

4.1. Research Design

In order to validate the annotation scheme and its guidelines, with particular emphasis on assessing consistency, reliability, and interpretability, we proceeded to Phase 3 of the study. This phase consisted of a small-scale pilot annotation conducted by two linguistics students familiar with the Text2Story scheme, who had extensive experience in referential annotation but no medical expertise. The annotation was subsequently curated by a linguist with a background in Pharmaceutical Sciences, who was also responsible for developing the Med2Story Referential scheme. The corpus selected for this phase comprised 15 expert-generated synthetic medical reports concerning

patients diagnosed with AML. The reports were authored by a Haematologist from IPO-Porto and corresponded to three fictitious patients, each represented by five reports: one group consultation report, three discharge reports, and one general report. The corpus contains a total of 1623 words; detailed corpus statistics are presented in Table 1. Annotation was performed using the INCEpTION platform (Klie et al., 2018).

Table 1: Summary statistics of report length across patient cases, including total word count (TWC), mean report length in words (MRL), and standard deviation (SD).

Patients	TWC	MRL	SD
Patient 1	611	122,8	13
Patient 2	494	98,8	24
Patient 3	518	103,6	16

Prior to the annotation task, a joint meeting was held involving the annotators and the curator. During this session, the guidelines were discussed, doubts were clarified, and one sample medical report was annotated collaboratively to ensure a shared understanding of the scheme. Upon completion of the annotation process, the corpus underwent curation, and the principal sources of disagreement were analysed, both between the annotators and between the annotators and the curator.

Subsequently, inter-annotator agreement (IAA) was calculated. High IAA scores typically indicate clear and operational guidelines, whereas low agreement may arise from a range of factors (Artstein, 2017; Basile et al., 2021; Bayerl and Paul, 2011), such as ambiguities or conceptual difficulties that warrant further refinement of the scheme.

After completing the annotation task, the annotators were asked to complete a questionnaire using five-point Likert scales (1–5). The questionnaire assessed several dimensions: the clarity of the definition of markables to be annotated; the effectiveness of the adaptation of Text2Story to the medical domain; the adequacy of the scheme for identifying and classifying grammatical and medical-domain information; the clarity of definitions and illustrative examples; the robustness of the proposal in resolving ambiguities; the extent to which the scheme provides tools to clearly distinguish events from participants; and the overall coherence and clarity of the framework. The items evaluated through the Likert scales are presented in Table 2.

4.2. Annotation Results

The IAA results for the annotation done in Phase 3 are presented in Table 3. IAA was evaluated using Krippendorff's alpha. In addition, agreement with the curator, who developed the annotation guidelines, was also calculated. The curator's annotations were used as a gold standard, enabling the assessment of how closely the annotators' decisions align with the gold annotation.

Annotator 1 (ANN1) and Annotator 2 (ANN2) agreed with each other on the identification of 135 participants, as well as on the assignment of the corresponding top-level type (e.g. Person, Institution and Unit, Biological Entity, etc.). The gold standard, that is, the curator's annotation, identified 174 participants. ANN1 showed a higher level of agreement with the Curator (CUR), identifying 160 of the 174 participants marked by the curator, whereas ANN2 identified 142 of these 174.

Cases involving discrepancies in the identification (or non-identification) of markables corresponded to two categories: missed markables and non-aligned markables (cf. Table 3). The former refers to cases in which annotators failed to identify a markable present in the gold standard. The latter includes instances in which a markable was identified by the annotator but did not fully match the gold standard span, i.e., the begin and/or end boundaries did not correspond exactly. One example concerns the annotation of *gengivorragias* ["gingival bleeding"] as *Anatomical and Physiological Location*, although this nominalisation denotes an event and therefore should not be annotated in this layer. Non-aligned markables involved cases such as *o aspirado medular mostrou 55% de blastos* ["the bone marrow aspirate showed 55% blasts"]. In such instances, the annotators marked *o aspirado medular* ["the bone marrow"] as a *Biological Entity* with the sub-attribute *Bio_Finding*, whereas only *medular* should have been annotated as *Biological Entity* with the second-level attribute *Tissue*. In this context, the noun phrase *aspirado medular* refers to a diagnostic procedure and therefore corresponds to an event. Only *medular* provides referential information, corresponding to "of the marrow", which denotes a tissue.

Among the markables and top-level types on which ANN1 and ANN2 agreed (135 cases), they also agreed on the second-level attributes (e.g. Patient, Cell, Bio_Finding, etc.) in 108 cases. Agreement on second-level attributes was higher between ANN1 and CUR (146/160) and between ANN2 and CUR (129/142).

Krippendorff's alpha for IAA measuring agreement on markable identification and top-level type assignment was 0.644 for ANN1/ANN2, 0.848 for ANN1/CUR, and 0.701 for ANN2/CUR. With regard to agreement on second-level attributes, the values were 0.771 (ANN1/ANN2), 0.901 (ANN1/CUR), and 0.894 (ANN2/CUR). Overall, these results may be considered robust, indicating that the guidelines were generally clear and could be reliably applied,

Table 2: Items included in the annotator questionnaire, rated on five-point Likert scales.

Parameters	Questions
Markables	How clear and consistent are the guidelines for defining markables?
Grammatical Domain	How effective are the guidelines for identifying and classifying participants at the grammatical level?
Medical Domain	How well-defined and comprehensive are the tags for the medical domain?
Definitions	How clear and detailed are the definitions of each tag?
Ambiguities	How effective is the approach to resolving ambiguities?
Coherence and Clarity	How coherent and clear are the provided guidelines?
Examples	How relevant and illustrative of the tags are the examples?
Text2Story	How effective is the adaptation of the Text2Story scheme for medical reports?
Events vs. Participants	The guidelines provide clear tests that allow events to be distinguished from participants?

Table 3: IAA between ANN1 and ANN2, between ANN1 and the CUR, and between ANN2 and the CUR.

Metrics	ANN1/ANN2	ANN1/CUR	ANN2/CUR
complete_markables	135	160	142
missed_markables	1	3	1
not_aligned_markables	56	20	46
complete_attributes	108	146	129
missed_class_attributes	27	14	13
krip_alpha_markable_total	0.644	0.848	0.701
krip_alpha_attributes	0.771	0.901	0.894

even by annotators without formal training in the medical domain.

One of the main sources of disagreement, which may account for the lower IAA values observed at the second-level attribute, concerns the identification of patient gender. Within the top-level type *Person*, attributes include *Patient*, *Donor*, *Physician*, among others. Within the attribute *Patient*, the sub-attribute *Male*/*Female* is specified. In several cases, gender was not explicitly stated but could be inferred from contextual information, as in *Doente de 74 anos com hiperplasia benigna da próstata* ["74-year-old patient with benign prostatic hyperplasia"] or *Doente de 47 anos, com antecedentes de fibromiomas uterinos* ["47-year-old patient with a history of uterine fibroids"]. Reference to the prostate allows one to infer that the patient is male, whereas reference to uterine fibroids implies a female patient. In some cases, this information only appeared in the patient's third report, making it impossible to infer the patient's sex from the previous reports. The annotators handled such cases differently: ANN1 assigned the sub-attribute female (or male) to subsequent occurrences of the patient once gender had been inferred, whereas ANN2 refrained from doing so.

One way to address this issue would be to resort to cross-document annotation, incorporating the *Objectal_LINK* (OLINK) relation of the *Identity* type in order to capture these cases of coreference. Thus, once the annotator identifies that the patient is female or male, they could mark the nearest markable corresponding to *Patient* and establish, through an *Identity* link, the required coreferential relation, thereby resolving the issue without the need to annotate all preceding markables.

Regarding the analysis of other IAA values for second-level attributes, high agreement was achieved for most attributes (e.g., ANN1/ANN2, ANN1/CUR, and ANN2/CUR all reached 1.000 for *Gene_Chromosome*). However, a low IAA value was observed for the attribute *Cell* between ANN1 and ANN2 (-0.023). The IAA between ANN1 and CUR was similarly low (-0.019), whereas the agreement between ANN2 and CUR was high (1.000). This pattern suggests that the disagreement may be attributable to a lack of expertise on the part of ANN1, while ANN2 encountered no comparable difficulties. Detailed IAA results for all attributes are available in the project's GitHub Repository.

With regard to the analysis of the Likert scale results, ANN1 assigned the maximum score to all parameters (5/5). ANN2 assigned the maximum score to the parameters related to the *grammatical domain*, *coherence and clarity*, *examples*, the adaptation of *Text2Story*, and the tests provided to facilitate the distinction between events and participants. ANN2 assigned a score of 4/5 to the parameters related to the definition of *markables*, noting that the guidelines were clear, with some areas for improvement; to the *medical domain*, stating that the tags are clear and well contextualised, with minor adjustments needed; and to the definitions and ambiguity resolution, remarking that good strategies are provided to handle ambiguities, with a few exceptions. ANN2 reported that the main difficulty concerned the annotation and classification of adjectival participants (e.g., *urinary retention*). To address this issue, we refined the definition of *markable* in the annotation manual and included additional examples. The Likert scale results for each annotator are available in the GitHub Repository.

The results of the Phase 3 curation revealed the identification of a total of 174 participants. With regard to the *Lexical_Head* category, the majority of participants correspond to nouns (152 occurrences) (*paciente com 75 anos* ["75-year-old patient"]), followed by adjectives (21 occurrences) (*hipertensão arterial* ["arterial hypertension"]) and a single occurrence of a pronoun (*iniciou vancomicina 1 g EV a 10/08/2018, que cumpriu durante 7 dias* ["vancomycin 1 g IV was initiated on 10/08/2018, which was administered for 7 days"]).

Table 4: Number of occurrences (Count) of top-level attributes of *Type*.

Top-level attributes of *Type*	**Count**
Biological Entity	51
Chemical Agent	36
Person	28
Anatomical and Physiological Location	24
Institution and Unit	24
Pathogenic Agent	7
Document	4

Regarding the *individuation* attribute, 121 instances received the value *individual* (*o FT3* ["the FLT3"]), 48 the value *mass* (*aspirado medular* ["the bone marrow aspirate"]), and 5 the value *set* (*o cariótipo* ["the karyotype"]).

Concerning the *involvement* attribute, the majority of occurrences (105) were classified with the value *1* (*a próstata* ["the prostate"]), 57 as *UND* (*tratamento com azacitina* ["treatment with azacitidine"], 10 as *>1* (*blastos* ["blasts'"]), and 2 as *all* (*hipertensão arterial* ["arterial hypertension"]).

With respect to the medical domain tags, the results for top-level attributes are summarised in Table 4. The most frequent top-level attribute was *Biological Entity* (51 occurrences), with *Gene_Chromosome* being the most represented *second-level attribute* within this class (33 occurrences) (*o FT3* ["the FLT3"]). By contrast, the least frequent attribute was *Document* (4 occurrences) (*passo o presente relatório a pedido do doente* ["I issue this report at the patient's request"]). The results for the second- and third-level attributes can be found in the GitHub Repository.

5. Conclusions and Future Work

In this work, our objectives were to assess the applicability of ISO 24617-9: 2019 to the annotation of medical reports in European Portuguese and to define and integrate an extension adapted to the medical domain of the referential layer of the Text2Story annotation scheme (based on ISO 24617-9: 2019, with adaptations for European Portuguese).

Our results show that ISO 24617-9: 2019 can be applied to medical reports; however, several adaptations are required, particularly with regard to the inclusion of domain-specific labels such as *Biological Entity*, *Pathogenic Agent*, and *Anatomical and Physiological Location*, among others.

With respect to OLINKs, these are not sufficient to capture the relevant information in this type of text, making it necessary to propose additional links, which we intend to address in future work. Future research will also involve applying the proposed scheme to a larger dataset of real medical reports.

6. Limitations

Regarding the limitations of the annotation scheme, the main issue we would like to highlight is the current absence of links, which would facilitate the capture of relevant information. The links proposed in ISO 24617-9: 2019 are not sufficient to represent the referential information present in medical reports, as they are too generic and not adapted to the clinical context. Therefore, future work will involve proposing a set of links specifically designed for this annotation layer.

Another aspect that may be viewed either as a limitation or as a strength concerns the presence of several second- and third-level attributes within the Type category. Although these attributes make the annotation process denser and more demanding for the annotators, they also enable a more fine-grained and informative representation of the data.

Finally, another limitation concerns the use of synthetic reports and the size of the dataset. Although these reports were produced by a medical specialist, they lack the variability and complexity typically found in real clinical narratives. Nevertheless, this approach was adopted in order to address the ethical constraints associated with the sharing and use of authentic medical reports. In addition, the dataset is relatively small, which may limit the generalisability of the findings. However, the purpose of this annotation was to validate the proposed annotation scheme, which will be implemented in a larger dataset in future work.

7. Acknowledgements

This work was supported by national funds through the Fundação para a Ciência e a Tecnologia (FCT), under PhD grant 2025.00440.BD. The authors also acknowledge support from the StorySense project (DOI: 10.54499/2022.09312.PTDC). We thank Cláudia Couto and Cecília Ortiz for their contribution to the annotation process.

8. Bibliographical References

Alan R. Aronson and Francois M. Lang. 2010. An overview of metamap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Ron Artstein. 2017. Inter-annotator agreement. In N. Ide and J. Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 297–313. Springer Netherlands.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo

Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. 2020. Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 60–73. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): Integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Hagit Borer. 2003. Exo-skeletal vs. endo-skeletal explanations. In John Moore and Maria Polinsky, editors, *The Nature of Explanation in Linguistic Theory*. CSLI and University of Chicago Press.

Ana Maria Brito and Fátima Oliveira. 1997. Nominalization, aspect and argument structure. In Isabel Faria, Gabriela Matos, Maria de Miguel, and Inês Duarte, editors, *Interfaces in Linguistic Theory*, pages 57–80. Porto, Portugal.

Leonardo Campillos-Llanos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. 2018. A french clinical corpus with comprehensive semantic annotations: Development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, 52(2):571–601.

Leonardo Campillos-Llanos, Ana Valverde-Mateos, and Adrián Capllonch-Carrión. 2025. Hybrid natural language processing tool for semantic annotation of medical texts in spanish. *BMC Bioinformatics*, 26(1):7.

Inês Cantante, Rute Rebouças, Purificação Silvano, António Leal, and Evelin Amorim. 2024. "el niño who sobreviveu": Anotação semântica do *Harry Potter* em inglês, português europeu e espanhol. In *Book of Abstracts of the 40th National Meeting of the Portuguese Association of Linguistics*, pages 78–79. University of the Azores.

Louise Deléger, Leonardo Campillos-Llanos, Anne-Laure Ligozat, and Aurélie Névéol. 2017. Design of an extensive information representation scheme for clinical narratives. *Journal of Biomedical Semantics*, 8:37.

Dina Demner-Fushman, Willie J. Rogers, and Alan R. Aronson. 2017. Metamap lite: An evaluation of a new java implementation of metamap. *Journal of the American Medical Informatics Association*, 24(4):841–844.

María C. Durango, Ever A. Torres-Silva, and Andrés Orozco-Duque. 2023. Named entity recognition in electronic health records: A methodological review. *Healthcare Informatics Research*, 29(4):286–300.

Ana Luísa Fernandes, Purificação Silvano, António Leal, Nuno Guimarães, Evelin Amorim, and Alípio Jorge. 2025a. Can ISO 24617-1 go clinical? extending a general-domain scheme to medical narratives. In *Proceedings of the 21st Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-21)*, pages 41–52, Düsseldorf, Germany. Association for Computational Linguistics.

Ana Luísa Fernandes, Purificação Silvano, António Leal, Nuno Guimarães, Rita Rb-Silva, Luís Filipe Cunha, and Alípio Jorge. 2025b. The incremental process of building an annotation scheme for clinical narratives in Portuguese: the contribution of human variation analysis. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 332–343, Vienna, Austria. Association for Computational Linguistics.

Ana Luísa Fernandes. 2025. Temporal structure of clinical narratives in portuguese. Master's thesis, University of Porto.

Dan Flickinger, Stephan Oepen, and Emily M. Bender. 2017. Sustainable development and refinement of complex linguistic annotations at scale. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 353–377. Springer.

Carol Friedman. 2000. A broad-coverage natural language processing system. In *Proceedings of the AMIA Symposium*, page 270. American Medical Informatics Association.

Ana González-Moreno, Aalberto Ramos-González, Israel González-Carrasco, et al. 2025. A clinical narrative corpus on nut allergy: annotation schema, guidelines and use case. *Scientific Data*, 12:173.

Jane Grimshaw. 1990. *Argument Structure*. MIT Press.

Ayoub Harnoune, Maryem Rhanoui, Mounia Mikram, Siham Yousfi, Zineb Elkaimbillah, and

Bouchra El Asri. 2021. Bert based clinical knowledge extraction for biomedical knowledge graph construction and analysis. *Computer Methods and Programs in Biomedicine Update*, 1:100042.

Eduard Hovy and Julia Lavid. 2010. Towards a "science" of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation Studies*, 22:13–36.

ISO. 2019. *ISO 24617-9: 2019, Language resource management – Semantic annotation framework – Part 9: Reference annotation framework (RAF)*. The International Organization for Standardization, Geneva. Project leaders: Laurent Romary and Kiyong Lee (SC 4/ WG 2 convenor).

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.

Srinivasa Rao Kundeti, J Vijayananda, Srikanth Mujjiga, and M Kalyan. 2016. Clinical named entity recognition: Challenges and opportunities. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1937–1945.

António Leal, Purificação Silvano, Zuo Qinren, Evelin Amorim, and Alípio Jorge. 2025. An annotation scheme for financial news in Portuguese. In *Proceedings of the 21st Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-21)*, pages 63–75, Düsseldorf, Germany. Association for Computational Linguistics.

António Leal, Purificação Silvano, Evelin Amorim, Inês Cantante, Fátima Silva, Alípio Jorge, and Ricardo Campos. 2022. The place of iso-space in text2story multilayer annotation scheme. In *Proceedings of the 18th Joint ACL–ISO Workshop on Interoperable Semantic Annotation within LREC 2022*, pages 61–70. European Language Resources Association.

Effi Levi and Shaul R. Shenhav. 2022. A decomposition-based approach for evaluating inter-annotator disagreement in narrative analysis. ArXiv preprint.

Antonio Miranda-Escalada, Eulàlia Farré, and Martin Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, CEUR Workshop Proceedings, pages 303–323.

James Pustejovsky, Harry Bunt, and Annie Zaenen. 2017. Designing annotation schemes: From theory to model. In N. Ide and J. Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 31–63. Springer.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*.

James Pustejovsky and A. Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, 42(5):950–966.

Isabelle Roy and Elena Soare. 2013. Event Related Nominals. In Gianina Iordachioaia, Isabelle Roy, and Kaori Takamine, editors, *Categorization and Category Change*, pages 123–152. Cambridge Scholars Publishing, Newcastle upon Tyne.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Stefan Schulz, Warren Del-Pinto, Lifeng Han, Markus Kreuzthaler, Sareh Aghaei, and Goran Nenadic. 2023. Towards principles of ontology-based annotation of clinical narratives. In *Proceedings of the 14th International Conference on Biomedical Ontologies (ICBO 2023)*, pages 36–47, Sweden. CEUR Workshop Proceedings. 14th International Conference on Biomedical Ontologies (ICBO 2023), 28 Aug–1 Sep 2023.

Purificação Silvano, Alípio Mário Jorge, António Leal, Evelin Amorim, Hugo Sousa, Inês Cantante, Ricardo Campos, and Sérgio Nunes. 2023. Text2story lusa annotated corpus. Dataset.

Purificação Silvano, António Leal, Fátima Silva, Inês Cantante, Fatima Oliveira, and Alípio Mario Jorge. 2021. Developing a multilayer semantic annotation scheme based on ISO standards for the visualization of a newswire corpus. In *Proceedings of the 17th Joint ACL - ISO*

Workshop on Interoperable Semantic Annotation, pages 1–13, Groningen, The Netherlands (online). Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Ergin Soysal, Jin Wang, Min Jiang, Yanshan Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2018. Clamp – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. De Groen, Bradley Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46(Suppl 0):S5–S12.

Jiarui Yao, Harry Hochheiser, WonJin Yoon, Eli Goldner, and Guergana Savova. 2024. Overview of the 2024 shared task on chemotherapy treatment timeline extraction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 557–569, Mexico City, Mexico. Association for Computational Linguistics.

Lin Yao, Hong Liu, Yi Liu, Xinxin Li, and Muhammad Waqas Anwar. 2015. Biomedical named entity recognition based on deep neural network. *International Journal of Hybrid Information Technology*, 8(8):279–288.