Method for Designing Semantic Annotation of Sepsis Signs in Clinical Text

Melissa Y. Yan[1], **Lise Tuset Gustad** [2,3], **Lise Husby Høvik**[3,4], **Øystein Nytrø**[1]
[1] Dept. of Computer Science, Norwegian University of Science and Technology
[2] Faculty of Nursing and Health Sciences, Nord University
[3] Dept. of Circulation and Medical Imaging, Norwegian University of Science and Technology
[4] Clinic of Anaesthesia and Intensive Care, St. Olavs Hospital, Trondheim University Hospital
{melissa.yan, lise.t.gustad, lise.hovik, nytroe}@ntnu.no

Abstract

Annotated clinical text corpora are essential for machine learning studies that model and predict care processes and disease progression. However, few studies describe the necessary experimental design of the annotation guideline and annotation phases. This makes replication, reuse, and adoption challenging.

Using clinical questions about sepsis, we designed a semantic annotation guideline to capture sepsis signs from clinical text. The clinical questions aid guideline design, application, and evaluation. Our method incrementally evaluates each change in the guideline by testing the resulting annotated corpus using clinical questions. Additionally, our method uses inter-annotator agreement to judge the annotator compliance and quality of the guideline. We show that the method, combined with controlled design increments, is simple and allows the development and measurable improvement of a purpose-built semantic annotation guideline. We believe that our approach is useful for incremental design of semantic annotation guidelines in general.

1 Introduction

Annotated clinical text corpora provide natural language processing (NLP) and machine learning (ML) studies the data necessary to find patterns, classify, and predict patient risk and disease progression. Compared to models that only utilize structured data from the electronic health record (EHR), many studies and reviews have shown that model performance can increase by incorporating unstructured clinical text (Soguero-Ruíz et al., 2016; Huddar et al., 2016; Culliton et al., 2017; Assale et al., 2019; Sheikhalishahi et al., 2019; Spasic and Nenadic, 2020).

Pre-existing annotated clinical corpora include the Medical Information Mart for Intensive Care (MIMIC-III) (Johnson et al., 2016), the Clinical E-Science Framework (CLEF) (Roberts et al., 2007), and the Informatics for Integrating Biology and the Bedside (i2b2) challenges and National NLP Clinical Challenges (n2c2) (Uzuner and Stubbs, 2015; Luo et al., 2020). However, studies utilizing pre-existing annotated corpora must limit their research questions to the specific purpose(s) for which the corpus was annotated. Otherwise, the annotations required to answer a research question might be missing or too general. Thus, many studies opt to develop their own annotated clinical corpus tailored to capture and extract the necessary information for their research (Yim et al., 2015; Rama et al., 2018; South et al., 2009; Oliveira et al., 2022).

Methods with lower requirements for supervision, such as information extraction, commonly use keyword search, rule-based algorithms, and ML to detect clinical cases. However, those methods might not consider the context of the clinical case (Ford et al., 2016). For example, different documented signs within a specific situation can describe a medical condition that is not named. Hence, medical expertise is necessary for making annotation judgments and capturing clinical knowledge within the text (Xia and Yetisgen-Yildiz, 2012). Retrieving domain-specific patient knowledge to ascertain or answer clinical questions includes extracting data, information, and knowledge. Data are attributes (e.g., names or dates), information gives meaning to data (e.g., location, cause, and time), and knowledge interprets information based on one's role and responsibility (e.g., clinical document's purpose and effect) (Gudea, 2005).

Making a quality annotated corpus is an iterative process that includes designing an annotation guideline, annotating text with the guideline, and refining the guideline based on inter-annotator agreement (IAA) (Roberts et al., 2009; Xia and Yetisgen-Yildiz, 2012; Deleger et al., 2012; Savkov et al., 2016; Oliveira et al., 2022). Although studies describe how annotated clinical corpora were made, few studies are explicit about the design process.

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 236–246

We believe that the acquisition and transformation of clinical questions about the patient cohort into corresponding corpus requirements for retrieving information from the actual text of the annotated corpus should drive the annotation process.

2 Related Work

This section provides an overview of studies that describe the design process leading to an annotation guideline and annotated clinical corpus. Studies that share their annotation challenges or offer improvements are also included.

The CLEF Corpus was semantically annotated to help develop and evaluate the CLEF information extraction system (Roberts et al., 2007, 2009). Free-text documents in the corpus are histopathology reports, imaging reports, and clinical narratives (i.e., discharge summaries, reports, case notes, audits, letters, or narratives to the general practitioner, consultant, referrer, or patient). Initially, templates for the documents using ontology-based entities and relationships were manually filled-in. However, the templates did not directly align with text, and ontology complexity made it time-consuming to fill templates. Thus, Roberts et al. (2009) iteratively developed an annotation guideline based on a simplified version of the original ontology and template definitions. Following established standard NLP annotation methodology (Boisen et al., 2000), 2 clinicians annotated 31 documents over 5 sessions, and a third annotator resolved disagreements (Roberts et al., 2009). However, due to workload and time constraints, resigning annotators could have impacted the corpus quality and size. Thus, Roberts et al. (2009) proposed solutions such as pre-annotated documents and a reduced annotation scope.

The i2b2 challenges have annotated corpora for various purposes. For example, in the i2b2 NLP challenge of extracting patient medication from discharge summaries, 79 annotators from 20 teams annotated 251 discharge summaries in a community annotation experiment (Uzuner et al., 2010). The annotation guideline was developed iteratively in 2 phases before the community annotation. For several iterations in phase 1, university students annotated discharge summaries that were measured for IAA and asked questions to aid revisions. This produced a guideline and 17 annotated discharge summaries for phase 2. Finally, during phase 2, teams annotated discharge summaries using the guideline and addressed inconsistencies within the 17 annotated discharge summaries to produce a refined guideline.

The i2b2 temporal relations corpus contains 310 discharge summaries annotated by 8 annotators (Sun et al., 2013). The annotation guideline was based on the TimeML event and temporal expression specification language (Pustejovsky et al., 2003) and the Temporal Histories of Your Medical Event (THYME) project annotation guidelines. The corpus development process included: a guideline development pilot study, data selection, pre-annotation, annotator training session, 2 annotators annotating pre-annotated documents, an adjudicator who resolved disagreements, and evaluation.

The 2014 i2b2/UTHealth de-identification corpus annotation guideline focuses on removing Protected Health Information (PHI) in longitudinal medical records for automatic de-identification system development (Stubbs and Uzuner, 2015). Introduced PHI subcategories enable downstream analyses to adjust the scope or focus on specific categories. Additionally, they compared parallel and serial annotation processes on pre-annotated and unannotated corpora and found that the process does not affect annotation quality (Stubbs and Uzuner, 2017).

Xia and Yetisgen-Yildiz (2012) utilized a variation of the typical annotation process for 3 different studies. Each study's corpus focused on a specific clinical report, such as radiology, chest x-ray, or intensive care unit reports. The process included: defining a study based on clinical needs, selecting data, gaining ethical approval, writing annotation guidelines, creating annotation tools, annotating, building a system with the corpus, and testing if the system meets clinical needs. Physicians were guideline designers and annotators, whereas NLP researchers provided technical support and built NLP systems with the corpora. Suggestions for improvement included more NLP researcher involvement, consideration for guideline granularity versus annotation time, marking rationale or evidence for a label, and estimating time commitment.

Deleger et al. (2012) developed their annotation guideline by building off a previous guideline. The rest of the methods were similar: defining annotation tasks, selecting data from stratified random sampling, and annotating with 2 annotators. During the annotation process, 2 annotators annotated the same documents, IAA was measured, and con-

sensus sessions were held to resolve disagreements and update the guideline. Using the same annotation process, they built gold standard corpora from clinical trial announcements, US Food and Drug Administration (FDA) drug labels, and EHR clinical notes. This included clinical notes such as discharge summaries, referrals, reports, and notes for consultations, procedures, plans, or progress.

Interested in capturing infections caused by central venous catheters, a nurse specializing in infection annotated 2 745 of 22 174 inspected notes (Røst et al., 2018). Before inspection, duplicate notes were removed. The guideline was a table containing events for annotation. Defined by computer scientists, nurses, and an NLP domain expert, the annotation labels formed a hierarchy starting with generalized events at the top level and more specific events below. They also provided information about data access restrictions to promote patient confidentiality and clinical record extraction. Record extraction included physician and nurse notes for admissions, care, plans, evaluations, transfers, and discharge summaries.

In this study, we focus on a method of incremental annotation guideline design by intertwining acquisition with testing of corpus requirements and corresponding annotation phases. This ensures that the guideline produces an annotated corpus that fulfills corpus requirements derived from clinical questions, even if the clinical questions are not answerable by the actual data. To the best of our knowledge, there lacks a study that describes this approach in detail.

3 Objective

This study aims to describe our method for designing a semantically annotated corpus for signs of sepsis by starting from clinical questions that formulate the corpus requirements. Hence, the main contributions are: (1) providing a detailed description of the guideline design process before annotation, (2) illustrating the systematic and iterative annotation process taken, and (3) discussing insights from the design and annotation process.

3.1 Clinical Problem

Sepsis leads to life-threatening multi-organ failure and is caused by a dysregulated host immune response to an infection (Singer et al., 2016). One infectious agent is the *Staphylococcus aureus* (*S. aureus*) bacteria found on skin that is known to cause serious bloodstream infections (BSIs). There is a known overlap between sepsis and BSI, as BSI is found in 30–58% of sepsis patients depending on which sepsis definition is used (Phua et al., 2013; Mellhammar et al., 2021). An estimated 7.6%–35% of *S. aureus* BSIs are related to peripheral intravenous catheters (PIVCs), and the presence of phlebitis can indicate infection via PIVC (Mermel, 2017). A PIVC is a medical device inserted into a vein for administering intravenous (IV) fluids, medication, and blood transfusions. Unfortunately, improperly managed PIVCs can become gateways that lead to phlebitis, BSI, or sepsis (Zhang et al., 2016).

Despite the high sepsis mortality rates and routine usage of PIVCs, both sepsis and PIVCs are poorly documented in clinical text and rarely available as structured data in the EHR (Rohde et al., 2013; Alexandrou et al., 2018). This makes it challenging for hospitals to perform retrospective systematic quality surveillance of PIVC-related BSIs to lower sepsis incidents. Additionally, the lack of explicit documentation inhibits the opportunities for clinicians to learn from and improve PIVC care practices to lower BSI and sepsis rates.

4 Original Adverse Event Dataset

We had access to 18 555 Norwegian adverse event (AE) reports extracted from a hospital's electronic incident reporting system (Yan et al., 2021). Extracted AE reports described procedural deviations, misunderstandings, resource needs, and risky patient behavior. Each report has structured data (i.e., identifier, registration date, reporting hospital unit, if the event is patient-related or security-related, event type, and event severity) and an unstructured free-text note.

5 Semantic Annotation Design Process

This section presents the semantic annotation design process leading up to the annotation process and guideline development. A summary can be found in Figure 1.

5.1 Clarify and Operationalize Clinical Questions to Form Corpus Requirements

Curious about PIVC-related BSI or phlebitis that can lead to sepsis and opportunities to improve patient care, nurses proposed the clinical question: "Is there a connection between PIVCs and BSIs or PIVCs and phlebitis at the hospital?" Thus, the

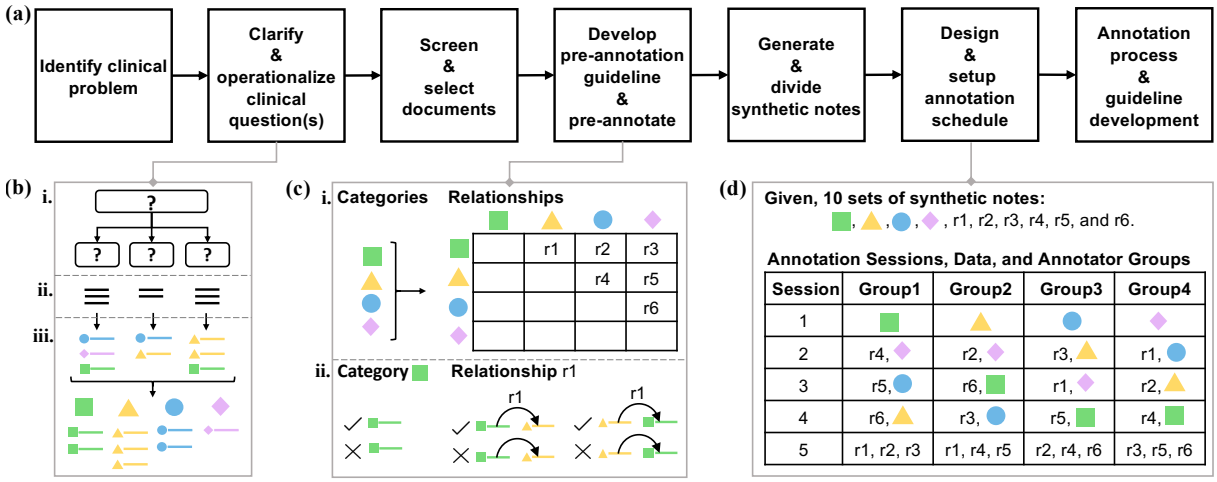

Figure 1: Semantic annotation design process. **(a)** Overview of the process until annotation and guideline development. **(b)** Clarify and operationalize clinical questions into corpus requirements to form annotation categories or entities. **i.** Clarify and operationalize clinical questions by expanding them to derive corpus requirements. **ii.** List examples to answer each question. **iii.** Sort examples into different categories to form the annotation categories. **(c)** Develop the pre-annotation guideline and pre-annotate. **i.** Find relationships using unique category combinations. **ii.** Create the pre-annotation guideline using concrete examples and counterexamples for categories and relationships. **(d)** Determine the annotation sessions and annotator groups to create a schedule. Divide synthetic notes into sets based on the number of categories and relationships. Each group annotates each category at least once in a different session. Additionally, each relationship is annotated at least twice by a different group throughout the sessions. Thus, the sets can be reused in different sessions by different groups, and guideline revisions can be tested on a different group using the same data.

clinical need is to identify PIVC-related BSI and phlebitis or sepsis signs, preferably by automatically classifying patients with PIVCs requiring follow-up care. Through iterative discussions with nurses and computer scientists, the clinical question was clarified to ensure data, information, and knowledge could be extracted to answer the clinical question (Figure 1 **(b)**i). Thus, the clinical question was clarified by expanding it into:

1. How can sepsis or BSIs be identified when the symptoms are similar to other diseases?
2. What signs or symptoms does PIVC-related phlebitis have?
3. How can poorly documented PIVCs be identified?

Those clinical questions were further modified based on the nurses' perspectives. For example, certain types of catheters are distinctly documented (for data extraction). Other catheters can be distinguished based on anatomical insertion sites (for information extraction) or procedures (for knowledge extraction). This resulted in the following questions that also operationalize and form the corpus requirements:

1. What are the different documented signs of infections or phlebitis, specifically those related to PIVCs, BSIs, or sepsis?
2. What can distinguish catheter types in the notes?
3. Where are the documented anatomical insertion sites of catheters?
4. What procedures, interventions, and activities can be related to catheter use from text content or report structured data (e.g., ward type or care situation)?

Figure 2 shows how clinical questions guide the design, application, and evaluation of the annotated corpus, annotation guideline, and corpus requirements.

Creating an annotated clinical corpus is time-consuming and labor-intensive (Wei et al., 2018). However, discussions revealed that we could not reuse a corpus and needed a new annotation guideline. Corpus requirements provided the annotation purpose and can be viewed as "information requests" to develop procedures for extracting data, information, and knowledge through annotation. Extracted data can be facts and observations, such

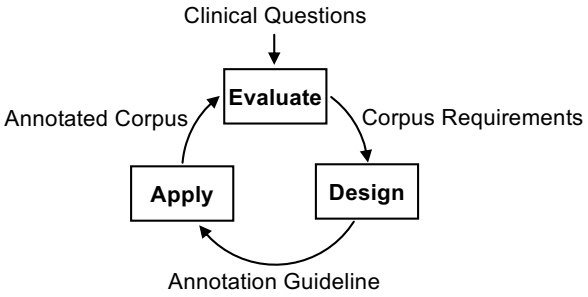

Figure 2: Design driven by clinical questions. Evaluating clinical questions forms the annotated corpus requirements used to design the annotation guideline. Annotators apply the annotation guideline to make a new (sub)corpus. The corpus is evaluated using inter-annotator agreement for annotator compliance and guideline comprehension. Clinical questions are used separately afterward to evaluate the corpus and requirements.

as dates, signs, or symptoms (e.g., purple skin). Information extracted can be phrases for specific signs and symptoms of a case (e.g., purple skin is a sign of a bruise). Furthermore, knowledge extracted can be other signs or symptoms that indicate something not necessarily mentioned (e.g., bruise color can indicate the stage).

Clarifying and operationalizing clinical questions helped determine corpus requirements about documented patient features, patient states, and care features. Including clinicians and computer scientists when clarifying questions was essential because it helped identify requirements for representing knowledge populated by text processing. Furthermore, these questions can be used to evaluate if the annotated corpus can answer the clinical questions.

5.2 Form Annotation Categories or Entities

Clinicians provided examples for the corpus requirements by listing keywords, phrases, and sentences (Figure 1 **(b)**ii). Computer scientists asked clarifying questions to resolve confusion and ambiguity. They also inquired about clinical actions versus actual documented actions to understand what is documented in the text. After generating a list of answers, answers were sorted into different categories (technically known as entities) through discussions (Figure 1 **(b)**iii). Each category is a label for a single word or phrase.

Answers were sorted into the 4 categories: **Sign**, **Location**, **Device**, and **Procedure**. Two additional categories, **Sensitivity** and **Person**, were included to ensure that data is de-identified and that the 4 categories can be linked to an individual. Thus, the 7 main categories are as follows:

1. **Sign**: infection signs
2. **Location**: anatomical insertion sites
3. **Device**: signs of catheter types
4. **Procedure**: catheter acts or interventions
5. **Sensitivity**: potential patient identifiers
6. **Person**: role (e.g., patient or clinician)
7. **Whole**: AE note topic label for validation (i.e., has patient identifier or is about infection, BSI, sepsis, faulty device, catheter, and/or PIVC).

Excluding the **Whole** category, the remaining 6 categories each form a hierarchy with more specific subcategories underneath. Subcategories are used to capture more detailed granularity from the text (e.g., the **Device** category contains a "Catheter" subcategory with different specific catheter types as subcategories).

Concrete categories made understanding the clinical annotation task easier and less ambiguous for the multidisciplinary research group. Having discussions and generating a list with clinicians helped determine the categories and subcategories needed to extract necessary data, information, and knowledge.

5.3 Screen and Select Notes

To ensure that categories specified above are present in notes, 700 randomly selected AE notes were manually screened and categorized by a computer scientist and nurse. Categorizing notes included providing a comment about the categorization rationale and marking potentially ambiguous notes. In addition, the potentially ambiguous notes were clarified in discussions and used as examples for properly annotating notes. Screening notes identified documented information that could satisfy corpus requirements and help answer clinical questions in downstream analyses. Additionally, it provided examples that drove preliminary guideline development in the next section.

5.4 Develop Pre-Annotation Guideline and Pre-Annotate

Initially, 6 possible relationships were found using a table with unique category combinations (Fig-

ure 1 **(c)**i). Then, those 6 relationships were discussed within the research group to evaluate which were required and merged. This resulted in the following 4 relationships for linking categories:

1. **Person** $\xrightarrow{\text{Person has}}$ **Sign**, **Location**, **Device**, or **Procedure**
2. **Procedure** $\xrightarrow{\text{Procedure uses}}$ **Device**
3. **Sign** $\xrightarrow{\text{Caused by}}$ **Device** or **Procedure**
4. **Sign**, **Device**, or **Procedure** $\xrightarrow{\text{Located nearby/on/in}}$ **Location**

Before actual annotation, the preliminary annotation guideline underwent a pre-annotation phase. Two pre-annotation guidelines were created to assess the utility and decide how detailed an annotation guideline should be for consistent annotation. The low granularity guideline was a Word document that provided brief instructions, a hierarchical list of categories, and only annotation examples for 2 categories (i.e., **Sensitivity** and **Person**). In contrast, the high granularity guideline was a static HTML webpage with interactive instructions for using the annotation tool and had links to corresponding sections for each category or relationship. Each category and the relationships in the high granularity guideline contained 1 concrete annotation example and counterexamples as needed (Figure 1 **(c)**ii). A nurse and a computer scientist used both pre-annotation guidelines to annotate 15-27 notes. Afterward, the research group determined a high granularity annotation guideline was more informative and easier to use with the annotation tool.

Capturing relationships between categories ensures that data is not lost in downstream analysis (e.g., infection signs at a specific location). It can also provide additional support to answer the clinical questions. By merging relationships, the complexity of annotation options was simplified and reduced. It is ideal to reduce the complexity of annotation because making the annotation task too difficult and time-consuming can result in annotators resigning (Roberts et al., 2009). The pre-annotation phase allowed the research group to manually evaluate, discuss, revise, and improve the guideline before use. This included the suitable granularity level and ease of use for the annotators.

5.5 Generate and Divide Synthetic Notes

Synthetic notes appear real and could be real. 100 unique synthetic clinical text notes were manually generated through 2 methods. The first method combines parts of the original notes to create a similar synthetic AE note with manually anonymized patient identifiers, and the content was verified by a nurse. Whereas in the second method, a nurse manually created a note based on possible clinical scenarios with synthetic patients to ensure some notes contained information about catheters and/or infections. The mean, minimum, maximum and median tokens per AE note in the corpus were 45, 4, 316, and 36, respectively. Generating synthetic notes took a couple of workdays for the nurse.

Afterward, the 100 unique synthetic notes were divided into 10 distinct sets with 10 notes each. Each set corresponds to either a category or relationship. The categories utilized in dividing the sets are those related to catheters or infections (i.e., **Sign**, **Location**, **Device**, and **Procedure**). The relationships utilized are the 6 initial possible relationship combinations.

AE notes often contain excessive and potentially identifying information irrelevant for annotating catheter-related events. Thus, relevant and closely related AE notes were selected and combined to use annotator time efficiently. Generating synthetic notes ensures the data is anonymized and usage is optimized, as clinical data is scarce. Additionally, it provides more data for ML analyses and makes the data more easily accessible to other researchers. Separating synthetic data into different categories or relationships ensures that specific labels will be annotated within the dataset. Different sets could be given to different annotators to reuse data and test if annotation guideline revisions improved IAA.

5.6 Design and Set Up Annotation Schedule

The same 4 categories and 6 initial relationships used to divide synthetic notes into 10 sets were used to design the annotation schedule (Figure 1 **(d)**). Categories were separated into groups, and relationships were added such that each group would annotate a relationship that excluded the group's category. Additionally, relationships within the groups were organized such that each relationship was annotated at least twice by 2 different annotator groups to evaluate revisions. This resulted in 4 annotator groups, each with 5 annotation sessions that

used a different set of notes and could annotate in parallel. Each group had 2 annotators so that IAA could be measured. This design defined the annotation schedule, the number of annotation groups needed, and how to reuse synthetic notes for guideline development. Furthermore, parallelization for each session helped reduce the project timeline.

6 Annotation Process and Annotation Guideline Development

Following the schedule, synthetic notes were annotated by 4 annotator groups over 5 sessions using a systematic, iterative annotation process for guideline refinement. In each session, 2 annotators from each group annotated notes based on an annotation guideline using the Brat rapid annotation tool (BRAT) (Stenetorp et al., 2012). Afterward, annotations were evaluated for IAA and manually inspected to assess if annotations could fulfill corpus requirements and answer the clinical questions. Text was tokenized and annotation labels were assigned to tokens before measuring the IAA F_1-score. Disagreements and ambiguities were discussed within the research group, and comments from annotators were incorporated. Next, a computer scientist revised the guideline based on discussions. Finally, the process was repeated with a new set of notes and the revised guideline. Figure 3 shows an example sentence annotated by 2 different annotators.

7 General Results from Sessions 1–5

Over 5 sessions, 8 annotators annotated 100 unique synthetic AE notes to produce 770 annotated synthetic AE notes. From session 1, it was clear that subcategory and attribute names should not be used in more than one category, and synonyms should be avoided. For example, simultaneously having "Name" as both a **Sensitivity** subcategory and an attribute for the **Person** category raised questions. Furthermore, annotators left relationships, attributes, and notes unannotated because they felt those notes were irrelevant to answering the clinical questions.

The need for annotating relationships, attributes, and all notes for ML was addressed in session 2. Red font emphasized guideline revisions, and the guideline began with an "Overview of Updated Instructions" section to aid annotators in identifying revisions. In sessions 2 and 3, the main revisions were correcting and including missing subcategories to address annotator concerns.

Session 4 provided a structured terminology for the guideline. A terminology was developed from the guideline to give structure and provide users quick insight into the annotated corpus for downstream analysis (Yan et al., 2023). This restructured the annotation guideline for session 5 by removing ambiguities and allowed AE note querying to answer the clinical questions. For example, the new **Observation** category encompasses the **Sign** category's signs and symptoms and the **Procedure** subcategory "Device malfunction signs." The computer scientist who revised the guidelines misinterpreted clinical knowledge and made incorrect assumptions in the previous sessions, so the terminology and restructured guideline were validated by nurses to ensure medical concepts were used correctly before session 5. The session 1–4 annotation guidelines were made available online[1] for Yan et al. (2021), and the session 5 annotation guideline was added online for this study. IAA for different sessions are in Figure 4.

8 Discussion

8.1 Design and Annotation Process

The annotation guideline development design process focuses on identifying the effect of the guideline on different categories, corpus content, and clinical questions. Categories were developed to answer different clinical questions and focus on localized guideline changes. Revising parts of specific category hierarchies made it possible to make controlled changes to specific subcategories in the annotation guideline and observe the impact on the annotated corpus, IAA, and clinical questions.

The annotation process greatly influences and drives guideline development. Clinical questions led to corpus requirements that developed the annotation guideline, which is applied on the annotated corpus and evaluated by the clinical questions. In turn, evaluating the annotated corpus also either indicates if it is possible to fulfill corpus requirements to answer clinical questions or detects a lack of corpus content needed for the clinical questions. Using the iterative process, we uncovered corpus requirements that the corpus content could not fulfill and could revise the requirements to drive guideline development and annotation.

[1] https://folk.ntnu.no/melissay/ae-guidelines/

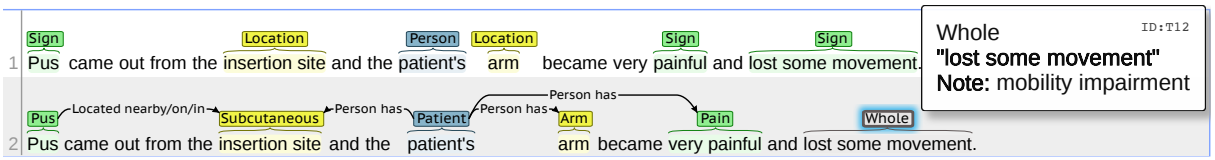

Figure 3: Annotation example for 2 different annotators. Annotator1 on top annotated using only the main categories, whereas Annotator2 on the bottom used subcategories to capture more detail and relationships to link categories. Although the **Whole** category is for indicating if an AE note contains information related to the clinical questions, Annotator2 has misused this label to leave a comment and indicate the phrase is about "mobility impairment". Actual AE notes only contain annotations from 1 annotator, and annotators cannot see the annotations from others.

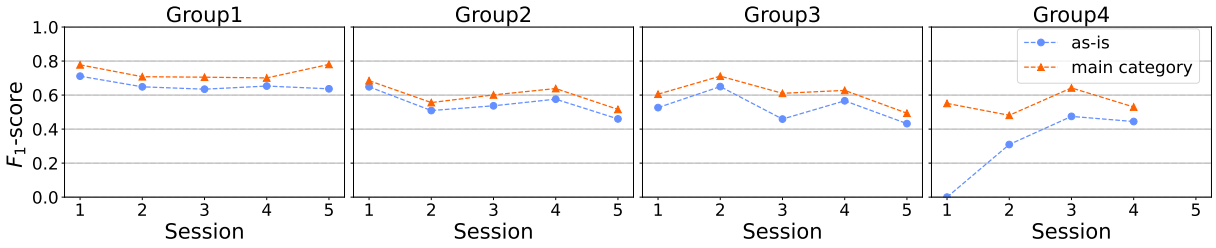

Figure 4: F_1-score for 4 annotator groups over the 5 sessions. The "as-is" F_1-score was calculated using annotator provided labels. Whereas, the "main category" F_1-score converted the labels to the main categories of **Sign**, **Location**, **Device**, **Procedure**, **Sensitivity**, **Person**, or **Whole**. Group4 session 5 has no F_1-score because an annotator withdrew.

8.2 Inter-annotator Agreement

There are several possible reasons for changes in Figure 4's F_1-score. An annotator often misused the **Whole** category to leave comments about clinical knowledge, while this is clinically insightful, it decreases the IAA (e.g., Figure 3 Annotator2 misusing the **Whole** category). The guideline complexity increased and reduced annotator compliance (i.e., sessions 1-5 had 89, 88, 105, 110, and 137 subcategories, respectively). As shown by the "as-is" F_1-score decrease in session 5, the guideline likely became too complex after session 4 revisions. The annotator from Group4 probably withdrew because of the increasing clinical complexity. Another annotator gave feedback that they were uncertain if they annotated some notes correctly. So, increasing the guideline and notes can overwhelm annotators (i.e., sessions 1-5 had 10, 20, 20, 20, and 30 notes, respectively). Group1 was a medical and nursing student, Group2 was a nurse and medical student, Group3 were nurses, and Group4 was a nurse and computer scientist. In general, students followed guidelines well, even if it contained incorrect medical concepts. Thus, paired annotators could have different clinical expertise that impacted results.

Granularity can have an effect on IAA, but granularity can be adjusted to identify problematic subcategories and utilized by those performing downstream analyses. Lower granularity in the annotation guideline leads to higher agreement because it reduces the complexity and level of detail. Annotators usually agree on which main category to annotate a word or phrase, but they had difficulties choosing certain subcategories. For example, in Figure 3 Annotator1 annotated with the main categories whereas Annotator2 was more detailed and annotated almost the same words with subcategories from the same main categories (e.g., "Pus" was annotated by Annotator1 with the **Sign** category and by Annotator2 with **Sign**'s subcategory "Pus"). This is also shown in Figure 4 for Group4 in session 1, where the "as-is" F_1-score is 0, but the "main category" F_1-score is 0.55. It is also possible to perform IAA on different subcategories within a subcategory to identify the most problematic areas after guideline revisions. The granularity in the annotated corpus can also be utilized and adjusted in downstream analyses based on the level of detail required by researchers.

9 Conclusion

Our method captures knowledge about sepsis signs in clinical text. We control changes in the annotation guideline by using hierarchical categories and continuous evaluation. Through applying a systematic, iterative annotation process, we evaluated the

changes using the clinical questions and IAA. The clinical questions evaluate corpus quality, and IAA evaluates annotator compliance and guideline complexity. As the guideline is designed to answer different clinical questions, it is possible to adjust the granularity level as needed to answer different clinical questions. By detailing our design process and annotation process, we hope our method can aid other researchers who cannot utilize pre-annotated corpora in developing an annotated corpus for their research.

Limitations

This method for designing and annotating clinical text for a specific clinical use case can be beneficial for researchers needing to annotate a corpus. However, there are some limitations. First, the experiences are based on a specific clinical case and focus on the qualitative aspects. Details of certain parts of the design and annotation process will likely need to be adjusted based on resources available to other researchers. This can include the data selected for annotation, the number of annotators available, and the annotators' level of expertise. For instance, the use case in the design process is based on using 8 annotators to annotate 100 synthetic AE notes over 5 sessions. Second, expertise and additional time are required to generate synthetic notes for annotation. Finally, future work is still needed to replicate the described design and annotation process on other forms of clinical text and problems.

Ethical Considerations

To protect patient privacy when designing and annotating clinical text, synthetic AE notes were manually generated and verified by a nurse to ensure the data is anonymized. Additionally, the annotation guideline includes the **Sensitivity** category to allow annotators to label potential information in the synthetic notes that could identify a patient. This process was described to provide an example for researchers who need to annotate sensitive data.

The Norwegian Regional Committees for Medical and Health Research Ethics (REK) has approved the use of medical data in this study (REK approval no. 26814; 2018/1201/REKmidt). To ensure annotators are protected, collecting and processing personal annotator data has also been approved by the Norwegian Centre for Research Data (NSD reference no. 142683). Furthermore, the annotators have consented to the use of their specified personal information (i.e., profession and years of experience) and their annotations.

Acknowledgements

The authors would like to thank the 8 annotators, this project would not be possible without them. Research is funded by the Computational Sepsis Mining and Modelling project through the Norwegian University of Science and Technology (NTNU) Health Strategic Area.

References

Evan Alexandrou, Gillian Ray-Barruel, Peter J. Carr, Steven A. Frost, Sheila Inwood, Niall Higgins, Frances Lin, Laura Alberto, Leonard Mermel, Claire M. Rickard, and O.M.G. Study Group. 2018. Use of short peripheral intravenous catheters: Characteristics, management, and outcomes worldwide. *Journal of hospital medicine*, 13.

Michela Assale, Linda Greta Dui, Andrea Cina, Andrea Seveso, and Federico Cabitza. 2019. The revival of the notes field: Leveraging the unstructured content in electronic health records. *Frontiers in Medicine*, 6:66.

Sean Boisen, Michael Crystal, Richard M. Schwartz, Rebecca Stone, and Ralph M. Weischedel. 2000. Annotating resources for information extraction. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000*, Athens, Greece. European Language Resources Association.

Phil Culliton, Michael Levinson, Alice Ehresman, Joshua Wherry, Jay S Steingrub, and Stephen I Gallant. 2017. Predicting severe sepsis using text from the electronic health record. In *Workshop on Machine Learning For Health at the Conference on Neural Information Processing Systems (NIPS ML4H 2017)*, Long Beach, CA, USA.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, and Imre Solti. 2012. Building gold standard corpora for medical natural language processing tasks. *AMIA Annual Symposium proceedings. AMIA Symposium.*, 2012:144–153.

Elizabeth Ford, John A. Carroll, Helen E. Smith, Donia Scott, and Jackie Cassell. 2016. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 23:1007 – 1015.

Sorin Gudea. 2005. Data, information, knowledge: a healthcare enterprise case study. *Perspectives in Health Information Management*, 2:8.

Vijay Huddar, Bapu Koundinya Desiraju, Vaibhav Rajan, Sakyajit Bhattacharya, Shourya Roy, and Chandan K. Reddy. 2016. Predicting complications in critical care using heterogeneous clinical data. *IEEE Access*, 4:7988–8001.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad M. Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.

Yen-Fu Luo, Sam Henry, Yanshan Wang, Feichen Shen, Ozlem Uzuner, and Anna Rumshisky. 2020. The 2019 n2c2/UMass Lowell shared task on clinical concept normalization. *Journal of the American Medical Informatics Association*, 27(10):1529–e1.

Lisa Mellhammar, Fredrik Kahn, Caroline Whitlow, Thomas Kander, Bertil Christensson, and Adam Linder. 2021. Bacteremic sepsis leads to higher mortality when adjusting for confounders with propensity score matching. *Scientific reports*, 11(1):6972.

Leonard A. Mermel. 2017. Short-term peripheral venous catheter-related bloodstream infections: A systematic review. *Clinical Infectious Diseases*, 65(10):1757–1762.

Lucas Emanuel Silva E Oliveira, Ana Carolina Peters, Adalniza Moura Pucca da Silva, Caroline Pilatti Gebeluca, Yohan Bonescki Gumiel, Lilian Mie Mukai Cintho, Deborah Ribeiro Carvalho, Sadid Al Hasan, and Claudia Maria Cabral Moro. 2022. SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. *Journal of Biomedical Semantics*, 13(1):13.

Jason Phua, Wang Ngerng, Kay See, Chee Tay, Timothy Kiong, Hui Lim, Mei Chew, Hwee Yip, Adeline Tan, Haji Khalizah, Rolando Capistrano, Kang Lee, and Amartya Mukhopadhyay. 2013. Characteristics and outcomes of culture-negative versus culture-positive severe sepsis. *Critical Care*, 17(5):R202.

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, Papers from 2003 AAAI Spring Symposium*, pages 28–34, Stanford University, Stanford, CA, USA. AAAI Press.

Taraka Rama, Pål Brekke, Øystein Nytrø, and Lilja Øvrelid. 2018. Iterative development of family history annotation guidelines using a synthetic corpus of clinical text. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 111–121, Brussels, Belgium. Association for Computational Linguistics.

Angus Roberts, Robert J. Gaizauskas, Mark Hepple, Neil Davis, George Demetriou, Yikun Guo, Jay Kola, Ian Roberts, Andrea Setzer, Archana Tapuria, and Bill Wheeldin. 2007. The CLEF corpus: Semantic annotation of clinical text. *AMIA Annual Symposium proceedings. AMIA Symposium*, pages 625–629.

Angus Roberts, Robert J. 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.

Jeffrey M. Rohde, Andrew J. Odden, Catherine Bonham, Latoya Kuhn, Preeti N. Malani, Lena M. Chen, Scott A. Flanders, and Theodore J. Iwashyna. 2013. The epidemiology of acute organ system dysfunction from severe sepsis outside of the intensive care unit. *Journal of hospital medicine*, 8:243–247.

Thomas Brox Røst, Christine Raaen Tvedt, Haldor Husby, Ingrid Andås Berg, and Øystein Nytrø. 2018. Capturing Central Venous Catheterization Events in Health Record Texts*. In *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, pages 488–495, Madrid, Spain. IEEE Computer Society.

Aleksandar Savkov, John A. Carroll, Rob Koeling, and Jackie Cassell. 2016. Annotating patient clinical records with syntactic chunks and named entities: the harvey corpus. *Language Resources and Evaluation*, 50:523 – 548.

Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani. 2019. Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Medical Informatics*, 7(2):e12239.

Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, and et al. 2016. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):801–810.

Cristina Soguero-Ruíz, Kristian Hindberg, I. Mora-Jiménez, José Luis Rojo-álvarez, Stein Olav Skrøvseth, Fred Godtliebsen, Kim Mortensen, Arthur Revhaug, Rolv-Ole Lindsetmo, Knut Magne Augestad, and Robert Jenssen. 2016. Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods. *Journal of Biomedical Informatics*, 61:87–96.

Brett R South, Shuying Shen, Makoto Jones, Jennifer Garvin, Matthew H Samore, Wendy W Chapman, and Adi V Gundlapalli. 2009. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. *BMC Bioinformatics*, 10 Suppl 9:S12.

Irena Spasic and Goran Nenadic. 2020. Clinical text data in machine learning: Systematic review. *JMIR Medical Informatics*, 8(3):e17984.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58 Suppl:S20–S29.

Amber Stubbs and Özlem Uzuner. 2017. De-identification of medical records through annotation. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 1433–1459. Springer Netherlands, Dordrecht.

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

Özlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. 2010. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5):519–523.

Özlem Uzuner and Amber Stubbs. 2015. Practical applications for natural language processing in clinical research: The 2014 i2b2/UTHealth shared tasks. *Journal of Biomedical Informatics*, 58 Suppl:S1–S5.

Qiang Wei, Amy Franklin, Trevor Cohen, and Hua Xu. 2018. Clinical text annotation - what factors are associated with the cost of time? *AMIA Annual Symposium proceedings. AMIA Symposium*, 2018:1552–1560.

Fei Xia and Meliha Yetisgen-Yildiz. 2012. Clinical corpus annotation: challenges and strategies. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in conjunction with the International Conference on Language Resources and Evaluation (LREC 2012)*, pages 32–39.

Melissa Y. Yan, Lise Tuset Gustad, Lise Husby Høvik, and Øystein Nytrø. 2023. Terminology and ontology development for semantic annotation: A use case on sepsis and adverse events. *Semantic Web*, 14(5):811–871.

Melissa Y. Yan, Lise Husby Høvik, André Pedersen, Lise Tuset Gustad, and Øystein Nytrø. 2021. Preliminary processing and analysis of an adverse event dataset for detecting sepsis-related events. In *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021, Houston, TX, USA, December 9-12, 2021*, pages 1605–1610. IEEE.

Wen-wai Yim, Sharon Kwan, and Meliha Yetisgen. 2015. In-depth annotation for patient level liver cancer staging. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 1–11, Lisbon, Portugal. Association for Computational Linguistics.

Li Zhang, Siyu Cao, Nicole Marsh, Gillian Ray-Barruel, Julie Flynn, Emily Larsen, and Claire M Rickard. 2016. Infection risks associated with peripheral vascular catheters. *Journal of Infection Prevention*, 17(5):207–213.