Medical Entity Corpus with PICO Elements and Sentiment Analysis

Markus Zlabinger[1], Linda Andersson[1], Allan Hanbury[1],
Michael Andersson[2], Vanessa Quasnik[3], Jon Brassey[4]
[1]TU Wien, [2]Stockholm University, [3]Uppsala University, [4]Trip Database
Austria, Sweden, United Kingdom
[1]forename.surname@tuwien.ac.at, [2]micche@gmail.com, [3]vanessa@quasnik.de, [4]jon.brassey@tripdatabase.com

Abstract

In this paper, we present our process to establish a PICO and a sentiment annotated corpus of clinical trial publications. PICO stands for Population, Intervention, Comparison and Outcome — these four classes can be used for more advanced and specific search queries. For example, a physician can determine how well a drug works only in the subgroup of children. Additionally to the PICO extraction, we conducted a sentiment annotation, where the sentiment refers to whether the conclusion of a trial was positive, negative or neutral. We created both corpora with the help of medical experts and non-experts as annotators.

Keywords: sentiment analysis, PICO, medical corpus, annotation

1. Introduction

Text mining, like data mining or knowledge discovery, is a process to discover implicit knowledge and potentially useful patterns from large text collections. For machine learning (ML) algorithms, it is essential that the labeled data is designed in such a way that the optimal learning for an algorithm is achieved.

We extracted PICO (Figure 1) elements from the text of publications of clinical trial results in order to improve a medical search mechanism of clinical questions. In this paper, we present the process of creating an annotated, phrase-level PICO (Population, Intervention, Comparison, Outcome) corpus and a sentence-level sentiment analysis corpus. In contrast, in other PICO annotation approaches (Boudin et al., 2010; Kim et al., 2011; Wallace et al., 2016), the PICO elements were only labeled on a sentence-level or abstract-level.

The PICO elements are associated with different aspects of noun phrases and domain terminology. The Population (P) elements generally consist of a patient description (e.g. children, men) with one or more post modifications; e.g., *patient over forty with type 2 diabetes*. The Intervention (I) respectively Comparison (C) describes a treatment method (e.g. drug treatment, surgery) and the Outcome (O) describes the aim of a conducted study (e.g. reduce pain).

Our contributions can be summarized in three main points: i) creation of a corpus dataset labeled with PICO elements on a phrase-level, ii) a sentiment analysis of the Outcome, i.e. if an Intervention had a positive effect on a target population or not, and iii) description of a new annotation methodology for community annotation of medical data, which requires less domain-specific knowledge.

The rest of the paper is organized as follows: In Section 2, existing PICO corpora and annotation schemata are presented. Section 3 contains a description of the data and the developed annotation interfaces. The conclusion and future work are described in Section 4.

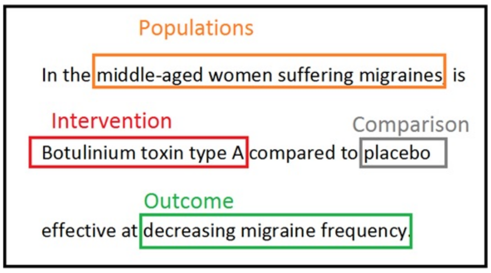

Populations
In the middle-aged women suffering migraines is
Intervention
Comparison
Botulinium toxin type A compared to placebo
Outcome
effective at decreasing migraine frequency.

Figure 1: Example of PICO elements in a sentence

2. Related Work

As mentioned, there are existing datasets based on PICO elements. For instance, (Huang et al., 2006) presented PICO frames as a knowledge representation by analysing 59 real-world, primary-care clinical questions. They obtained the questions from the Family Practice Inquiries Network (FPIN) and Parkhurst Exchange. (Demner-Fushman and Lin, 2007) annotated PICO elements in 275 PubMed abstracts and used the data to build a clinical question-answering system.

(Kim et al., 2011) established a corpus of 1,000 medical abstracts, which were manually annotated with 6 different categories (Background, Population, Intervention, Outcome, Study Design, Others) on a sentence-level by medical experts. (Boudin et al., 2010) extracted 260,000 abstracts from PubMed. They limited their domain to English abstracts with publication dates from 1999 to 2009; and used the publication categories Humans, Clinical Trial (CT) and Randomized Controlled Trial (RCT). They exploited the sentence headings that occur in some abstracts (e.g. Results, Methods, . . .). Some of these headings are an indicator for certain PICO elements; for example, the headings *Participants* or *Sample* indicate a Population. Based on the sentences that occur within certain headings, they created a dataset and evaluated several ML classifiers on it. In (Wallace et al., 2016), they exploited an already existing semi-structured resource, the Cochrane Database of Systematic Reviews (CDSR). They derived supervised distant supervision from the CDSR resource in order to obtain data to train a PICO extraction model. In a follow-up study (Marshall et al., 2017), they used a modified version of this semi-automatic annotated dataset to create a prototype to extract and synthesize medical evidence information from clinical trial articles.

In all of the above studies, domain experts were involved in the data creation process. An expert annotation task is defined as a task that requires several years of knowledge or a specific profession in order to understand and conduct the task correctly (Xia and Yetisgen-Yildiz, 2012). Expert annotators are rather expensive and given the amount of data required to train the end-applications, are a bottleneck in the development of domain-specific text mining applications.

Moreover, the expert annotation schema is not a guarantee for high quality annotations, since domain knowledge alone is not enough to annotate sought entities. The technical aspect of annotation also requires an understanding of what an algorithm can learn from the labeled data (Xia and Yetisgen-Yildiz, 2012). A well designed annotation task requires expert knowledge in the text domain as well as in fields related to linguistics, computational linguistics and Natural Language Processing (NLP), i.e. when annotating an entity, it is important to consider what an algorithm can learn from it. The technical gap between knowledge of the text domain and the requirement of the algorithms to learn the sought entities has been a topic of several publications (Uzuner et al., 2010; Yetisgen-Yildiz et al., 2010; Xia and Yetisgen-Yildiz, 2012). Due to the technical gap between domain experts and artificial intelligence experts, the annotation design for domain specific text genres generally are time consuming, expensive and the outcome is uncertain.

There are also less expensive annotation schemata, such as Crowdsourcing or Community Annotation. These could be alternatives if the domain specific annotation task is designed well. The Crowdsourcing schema makes use of the online labour via annotation providers such as Amazon's Mechanical Turk or Crowd Flower. Depending on the task, these schemata can obtain good results at a low cost (Yetisgen-Yildiz et al., 2010). In the Community Annotation schema, annotations are gathered from the research community that is interested in a particular task and thereby have some pre-knowledge of the target domain and task, which can be beneficial (Uzuner et al., 2010).

3. The Annotation Process

As mentioned in Section 1, the final goal was to extract the PICO elements and the sentiment from medical publications. Since there was no appropriate data available to create an automatic PICO approach or sentiment classifier, we created it ourselves with the help of expert and non-expert annotators. The PICO annotation task requires a different level of linguistic information and also information from different medical domains, which further increases the complexity of the annotation task. For example, the decision of whether labeling an element as Population or Intervention depends on the context; i.e., in a different context, a drug can be part of a Population and in another context, part of an Intervention (see Table 1 for examples).

In addition, to the PICO annotation, we included a sentiment analysis annotation for a subset of the PICO labeled data. For the sentiment analysis, we defined three classes: positive, negative and neutral. The class selection depends on the outcome of the Intervention compared to the Comparison. That is, if an Intervention is better (e.g. more effective, less adverse effects) than its Comparison, the sentiment is positive. On the other hand, an Intervention that did not perform better than its Comparison, should be classified as negative. All other cases are classified as neutral.

We developed two annotation interfaces (one for the sentiment and one for the PICO) and tested them with a group of 6 persons who come from different backgrounds: linguists, biologists, medical experts and students. Step by step, we updated the interface to create a more effective annotation environment for the annotators. The main goal of these updates was to improve the agreement between the annotators; because, if not even humans can agree on where the PICO elements are located or what sentiment a publication has, an algorithmic approach will most certainly also fail to do so. In addition, better agreements mean that the resulting dataset is more reliable and therefore it will be easier to create a well performing automatic detection approach.

3.1. Data Collection

We were provided with 1.5M PubMed titles and abstracts from Trip[1] of which a subsample was used in this first attempt to establish a PICO and sentiment corpus. Since not all publication types are of interest for sentiment analysis and PICO extraction, we used exclusively Randomized Control Trials (RCTs), which contain the following key components: an intervention-arm (*aspirin*), a comparison-arm (*placebo*), an outcome (*Aspirin is more effective than placebo*) and finally, a group of people who are randomly assigned to the intervention-arm or comparison-arm (*men with headache were randomized to either [...]*).

3.2. Annotation Infrastructure

All interfaces for the sentiment and the PICO annotation were implemented by using a mixture of HTML5, JavaScript and PHP5. The submitted annotations were saved in a MySQL database. The publications (i.e. abstract, title) and the user information (i.e. username, user id, etc.) were also stored in the MySQL database.

To increase the agreement between annotators, we provided two guideline documents: one for the PICO annotation task and one for the sentiment annotation task. The guidelines can be seen as a reference manual that can be referred to for difficult cases, but also as an introduction on how to accurately identify the important text parts that should be annotated. The guidelines were updated based on the annotations that we got from the users in small-scale test runs.

3.3. The PICO Annotation Tool

In this section, the different prototypes of the PICO interface are presented. In total, we created three versions of the PICO annotation tool. For each version, the agreement

[1] https://www.tripdatabase.com/

Example	Population	Intervention
Adverse effects of aspirin in men who take vitamin C regularly	men who take vitamin C regularly	aspirin
Adverse effects of vitamin C in men	men	vitamin C
Effects of paracetamol in patients who underwent bankart repair	patients who underwent bankart repair	paracetamol
Bankart repair in patients with shoulder instability	patients with shoulder instability	Bankart repair

Table 1: Depending on the context, treatment methods can be part of a Population or an Intervention.

between annotators was computed. The aim was to successively reach better agreements after each interface update.

3.3.1. First Prototype (version 0)

In the first version of the annotation tool (see Figure 2), the user was asked to mark text within the title or the abstract. Then the marked text could be assigned to one of the four PICO classes with a single button click. Additionally, we allowed open text input for cases where the PICO information was only implicitly stated; for example, *placebo-controlled trial* would mean that the Comparison is placebo. We also offered an advice system that was based on rules crafted with Stanford's TokensRegex (Chang and Manning, 2014). TokensRegex is a rule based framework and used for information extraction. It is similar to regular expressions, but applied over NLP components (e.g. part of speech tags, word tokens) rather than single characters.

3.3.2. Second Prototype (version 1)

Since the open text input from version 0 lead to low inter-annotation agreements of about 20%, we developed a more restricted interface with respect to user interaction for the second prototype. In this version, we first split the publication text into sentences and then each sentence into its tokens. To do so, we used Stanford's CoreNLP (Manning et al., 2014), which is an NLP toolkit that includes sentence splitting and tokenization. To give additional guidance, we provided semantic labels for some of the tokens; for example, diseases, drugs or persons were labeled, as illustrated in Figure 3. The semantic labels for the medical information were generated by using GATE's BioYodie pipeline (Wu et al., 2018), which is a tool for Named Entity Recognition in medical documents. To label Person elements, we simply used a static lookup list that consisted of 44 person keywords (e.g. patients, seniors, children).

In order to do an annotation in the second prototype, the annotator selects one sentence and afterwards, he/she selects the start and end token of the PICO element by simply clicking on them, i.e. open text input was prohibited in this version's interface. Afterwards, a pop-up window opens where the PICO type is selected (see Figure 3). Finally, the annotator selects one of the four PICO classes.

3.3.3. Third Prototype (version 2)

In the third prototype (final version), two changes were made: First, we decided to drop the Outcome from the PICO annotation task since it appeared to be too diverse to reach a reasonable inter-annotator agreement, i.e. only PIC was annotated. Second, we introduced a confidence selection where the annotator could state how confident he/she was, in his/her annotation. We offered three options: Low Confidence, Medium Confidence and High Confidence (default). With the third prototype, we achieved acceptable agreements of around 45% for the Intervention/Comparison, and 55% for the Population, in a small-scale test run. The third prototype is illustrated in Figure 3, which is, besides the two mentioned changes, identical to the previously described interface (i.e. version 1).

We decided to establish the first version of the dataset using majority voting; e.g., if two or more annotators labeled the same part of a text as Population, we considered it as a Population annotation. Based on this majority voting strategy, we started the final annotation run with our six annotators. We distributed 50 unique documents to each annotator and then 50 community documents, which were identical for all 6 annotators. This document distribution was repeated until a total of 500 documents were assigned to each annotators' account. To sum it up, each annotator annotated 250 unique and 250 community documents, which makes a total of $(6 \times 250) + 250 = 1750$ annotated RCTs.

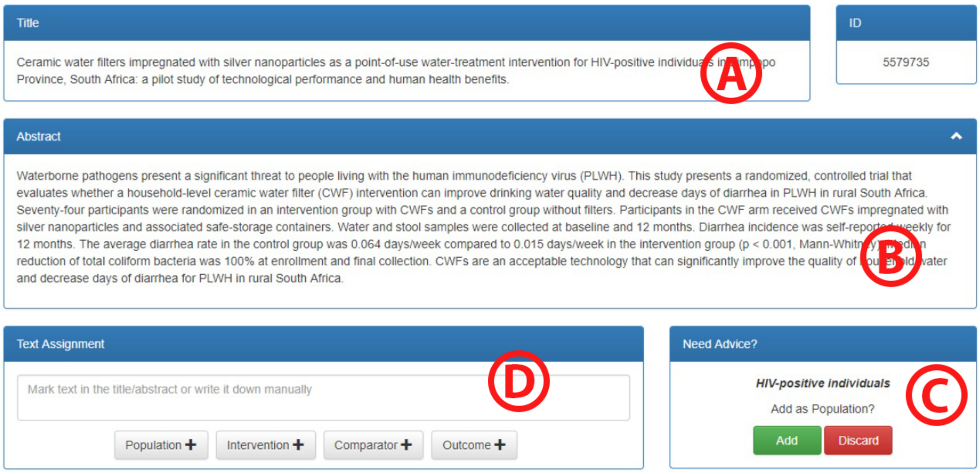
We observed from the annotated dataset that the experts had a tendency to add the design (e.g. randomized, blind) of the trial to either the Population or Intervention. For the Population, they occasionally forgot to mark the entire noun phrase. Meanwhile, the non-experts had difficulties in identifying Populations that had no reference to a Person entity (e.g. *apsirin in headache* VS *aspirin in men with headache*).

3.4. The Sentiment Annotation Tool

For a subset of the PIC annotated corpus, the sentiment was annotated. We differentiated between two types of RCTs:

- Type 1: The abstract contains a conclusion section, as is the case for the abstract shown in Figure 4. In this case, we asked the annotator to select a sentiment of either positive, neutral (default) or negative. Afterwards, by clicking submit, the sentiment annotation is saved in the database.
- Type 2: The abstract does not contain a conclusion section, as is the case for the abstract shown in Figure 5. In this case, we asked the annotators to click on the first sentence where he/she thinks that the conclusion starts. This clicked sentence and all subsequent ones were then listed and a sentiment could be selected for each one.

With the developed interface, it was possible to achieve an inter-annotator agreement of 80%. Note: Before computation of the agreement, the negative and neutral classes were merged, since negative sentiments occurred too rarely (in $\sim 10\%$ of the cases).



Ceramic water filters impregnated with silver nanoparticles as a point-of-use water-treatment intervention for HIV-positive individuals in Limpopo Province, South Africa: a pilot study of technological performance and human health benefits.

5579735

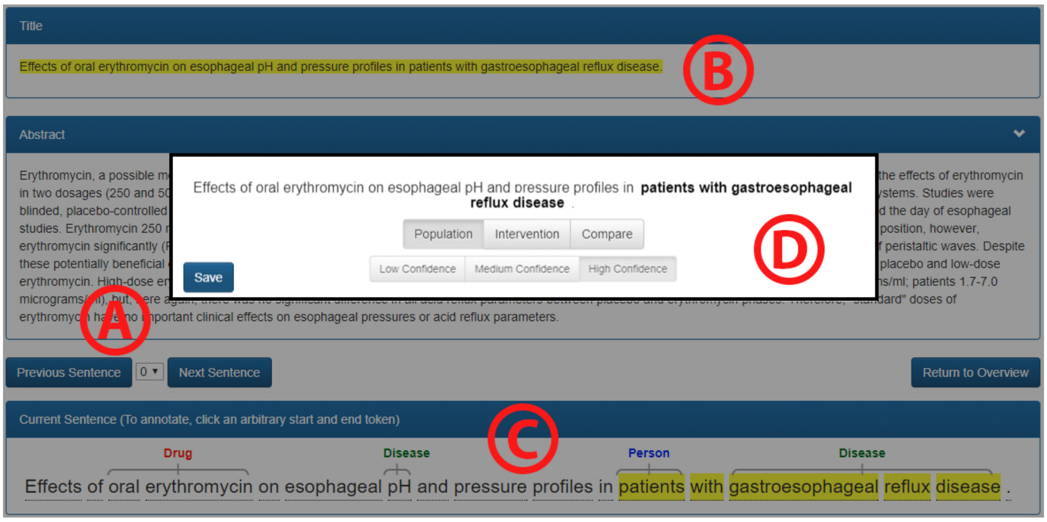
Waterborne pathogens present a significant threat to people living with the human immunodeficiency virus (PLWH). This study presents a randomized, controlled trial that evaluates whether a household-level ceramic water filter (CWF) intervention can improve drinking water quality and decrease days of diarrhea in PLWH in rural South Africa. Seventy-four participants were randomized in an intervention group with CWFs and a control group without filters. Participants in the CWF arm received CWFs impregnated with silver nanoparticles and associated safe-storage containers. Water and stool samples were collected at baseline and 12 months. Diarrhea incidence was self-reported weekly for 12 months. The average diarrhea rate in the control group was 0.064 days/week compared to 0.015 days/week in the intervention group (p < 0.001, Mann-Whitney). Median reduction of total coliform bacteria was 100% at enrollment and final collection. CWFs are an acceptable technology that can significantly improve the quality of household water and decrease days of diarrhea for PLWH in rural South Africa.

Mark text in the title/abstract or write it down manually

HIV-positive individuals

Add as Population?

Figure 2: First version of the PICO annotation tool: (A) Title, (B) abstract, (C) advice system and (D) open text input.



Effects of oral erythromycin on esophageal pH and pressure profiles in patients with gastroesophageal reflux disease.

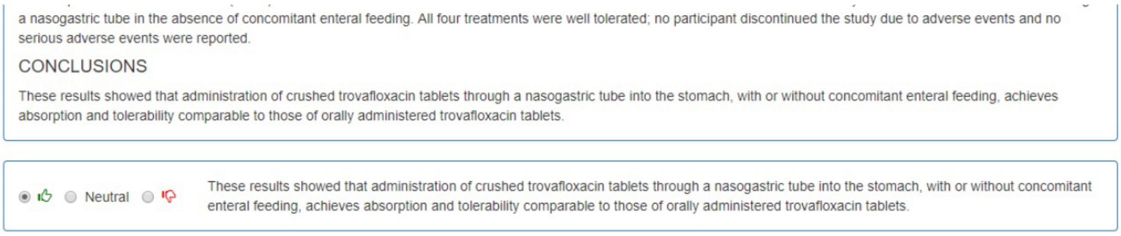
Erythromycin, a possible m ... the effects of erythromycin
in two dosages (250 and 50 ... ystems. Studies were
blinded, placebo-controlled ... d the day of esophageal
studies. Erythromycin 250 ... position, however,
erythromycin significantly (P ... f peristaltic waves. Despite
these potentially beneficial ... placebo and low-dose
erythromycin. High-dose er ... ns/ml; patients 1.7-7.0
micrograms/ml), but, here again, there was no significant difference in all acid reflux parameters between placebo and erythromycin phases. Therefore, "standard" doses of
erythromycin have no important clinical effects on esophageal pressures or acid reflux parameters.

Effects of oral erythromycin on esophageal pH and pressure profiles in **patients with gastroesophageal reflux disease** .

Current Sentence (To annotate, click an arbitrary start and end token)

Effects of oral erythromycin on esophageal pH and pressure profiles in patients with gastroesophageal reflux disease .

Figure 3: Final version of the PICO annotation tool: (A) Sentence navigation, (B) active sentence (yellow background), (C) active sentence split into single word units (tokens) and finally, after selecting a start and end token, a pop-up window (D) is shown and used to submit an annotation for either P, I or C.



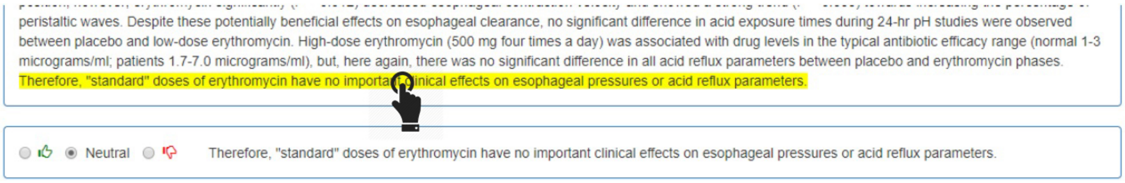
a nasogastric tube in the absence of concomitant enteral feeding. All four treatments were well tolerated; no participant discontinued the study due to adverse events and no serious adverse events were reported.

CONCLUSIONS

These results showed that administration of crushed trovafloxacin tablets through a nasogastric tube into the stomach, with or without concomitant enteral feeding, achieves absorption and tolerability comparable to those of orally administered trovafloxacin tablets.

● 👍 ○ Neutral ○ 👎 These results showed that administration of crushed trovafloxacin tablets through a nasogastric tube into the stomach, with or without concomitant enteral feeding, achieves absorption and tolerability comparable to those of orally administered trovafloxacin tablets.

Figure 4: (Type 1) The conclusion sentence(s) are shown immediately.



peristaltic waves. Despite these potentially beneficial effects on esophageal clearance, no significant difference in acid exposure times during 24-hr pH studies were observed between placebo and low-dose erythromycin. High-dose erythromycin (500 mg four times a day) was associated with drug levels in the typical antibiotic efficacy range (normal 1-3 micrograms/ml; patients 1.7-7.0 micrograms/ml), but, here again, there was no significant difference in all acid reflux parameters between placebo and erythromycin phases. Therefore, "standard" doses of erythromycin have no important clinical effects on esophageal pressures or acid reflux parameters.

○ 👍 ● Neutral ○ 👎 Therefore, "standard" doses of erythromycin have no important clinical effects on esophageal pressures or acid reflux parameters.

Figure 5: (Type 2) The starting sentence of the conclusion is selected by the annotator.

Since we already reached reasonable agreements in the first version of the annotation tool, we started the final annotation run in which we distributed 200 community and 200 unique documents to each annotator.

4. Conclusion

We have presented the process of establishing a PIC annotated corpus on a phrase-level. We collected a total of 1750 annotated RCTs (250 overlapping) by the annotations of both experts and non-experts. We also labeled a smaller set of these RCTs (1,400) with a sentiment. From our first version of the annotation interface, we increased the annotation agreement from 20% to 55% for the PIC elements. For the sentiment annotation analysis, we reached agreements of 80%.

We have developed an annotation tool for PIC and sentiment analysis, ready to be used in community annotation tasks. Furthermore, we discovered that the PIC annotation task can be conducted by non-experts, if the data is prelabeled with semantic categories, such as persons, drugs or diseases. We only observed minor annotation differences between non-experts and experts. The next step is to turn these two annotation tasks into a community annotation effort in order to collect more annotated data. As soon as we have increased the data, we plan to release part of the corpus to the research community.

5. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644753 (KConnect).

6. Bibliographical References

Boudin, F., Nie, J.-Y., Bartlett, J. C., Grad, R., Pluye, P., and Dawes, M. (2010). Combining classifiers for robust PICO element detection. *BMC Medical Informatics and Decision Making*, 10(1):29.

Chang, A. X. and Manning, C. D. (2014). TokensRegex: Defining cascaded regular expressions over tokens. Technical Report CSTR 2014-02, Department of Computer Science, Stanford University.

Demner-Fushman, D. and Lin, J. (2007). Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103, March.

Huang, X., Lin, J., and Demner-Fushman, D. (2006). PICO as a knowledge representation for clinical questions. In *AMIA 2006 Symposium Proceedings*, pages 359–363.

Kim, S. N., Martinez, D., Cavedon, L., and Yencken, L. (2011). Automatic classification of sentences to support evidence based medicine. *BMC Bioinformatics*, 12(2):S5.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Marshall, I., Kuiper, J., Banner, E., and Wallace, B. (2017). Automating biomedical evidence synthesis: Robotreviewer. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2.

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

Wallace, B. C., Kuiper, J., Sharma, A., Zhu, M. B., and Marshall, I. J. (2016). Extracting PICO sentences from clinical trial reports using supervised distant supervision. *Journal of Machine Learning Research*, 17(132):1–25.

Wu, H., Toti, G., Morley, K. I., Ibrahim, Z. M., Folarin, A., Jackson, R., Kartoglu, I., Agrawal, A., Stringer, C., Gale, D., Gorrell, G., Roberts, A., Broadbent, M., Stewart, R., and Dobson, R. J. (2018). SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *Journal of the American Medical Informatics Association*.

Xia, F. and Yetisgen-Yildiz, M. (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), Istanbul, Turkey*.

Yetisgen-Yildiz, M., Solti, I., Xia, F., and Halgrim, S. R. (2010). Preliminary experience with Amazon's mechanical turk for annotating medical named entities. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 180–183. Association for Computational Linguistics.