Annotation of pain and anesthesia events for surgery-related processes and outcomes extraction

Wen-wai Yim
Palo Alto Veterans Affairs
Stanford University
Palo Alto, CA 94305, USA
wwyim@stanford.edu

Dario Tedesco
University of Bologna
Bologna, 40126, Italy
Stanford University
Palo Alto, CA 94305, USA
dariot@stanford.edu

Catherine Curtin
Palo Alto Veterans Affairs
Stanford University
Palo Alto, CA 94305, USA
ccurtin@stanford.edu

Tina Hernandez-Boussard
Stanford University
Palo Alto, CA 94305, USA
boussard@stanford.edu

Abstract

Pain and anesthesia information are crucial elements to identifying surgery-related processes and outcomes. However pain is not consistently recorded in the electronic medical record. Even when recorded, the rich complex granularity of the pain experience may be lost. Similarly, anesthesia information is recorded using local electronic collection systems; though the accuracy and completeness of the information is unknown. We propose an annotation schema to capture pain, pain management, and anesthesia event information.

1 Introduction

Post surgical pain continues to be a challenging problem for the health system. Firstly, continued pain after surgery, or chronic persistent postsurgical pain, is common with about 20% of patients having pain long after the wounds have healed (Neil and Macrae, 2009; Kehlet et al., 2006). Secondly, inadequate acute post operative pain control contributes to adverse events such as impaired pulmonary function and impaired immune function (White and Kehlet, 2010). Finally, post surgical pain can be a gateway to addiction, which has taken on increased urgency with the current opioid crisis (Waljee et al., 2017). To improve these problems, it is crucial to have a clear understanding of the patients' pain and its treatments.

There is some evidence that different interventions such as the use of multi-modal pain management and different anesthesia types, e.g. use of regional anesthesia and nonsteroidal anti-inflammatory drugs, can improve pain management (Baratta et al., 2014). However, different analgesic treatments have different side-effect profiles; moreover, some treatment combinations are not appropriate for certain populations. Furthermore, genetics, age, prior exposure to surgery, and social norms influences the experience of pain. Therefore, there is a clear need to capture anesthesia and pain information and relate them to individual history, social, and genetic factors to improve surgical outcomes.

Even with mandated collection, pain is not always recorded (Lorenz et al., 2009). Even when recorded as structured data, there are a variety of scales that are institution-dependent, e.g. a site-specific 0-10 numeric rating scale or a multi-dimensional questionnaire such as the Brief Pain Inventory. Additionally, it is difficult to capture the rich complex characteristics of pain in structured ways. Anesthesia type, on the other hand, may be recorded or inferred from procedures, medications, or structured input as part of surgery documentation. However, such recording practices differ by institution and local software.

In this work, we present annotation schemas for pain, pain treatment, and anesthesia events for text extraction, as well as report on inter-annotator agreement and corpus statistics. The ultimate goal is to build a new system or adapt an existing system, using this annotated corpus, to automatically extract such information from clinical free text. The extracted data could then be used to complement missing structured information, facilitating greater opportunities for longitudinal study of patients' pain experience long after initial surgery.

Proceedings of the BioNLP 2017 workshop, pages 200–205,
Vancouver, Canada, August 4, 2017. ©2017 Association for Computational Linguistics

2 Related work

To our knowledge, there is no systematic creation of a pain annotation schema for text extraction, however we reference two extraction systems that identify pain information based on their own targeted needs. (Heintzelman et al., 2013) created a system that extracted pain mentions, severity, start date, end date. Their annotation was based on a created 4-value severity of pain created by the development team. Items were identified using the Unified Medical Language System (UMLS) vocabularies for dictionary look-up (Bodenreider, 2004). Dates and locations were extracted by developed contextual rules. In another work, (Redd et al., 2016) used a series of regular expressions to extract pain score in intensive care unit notes. In contrast to previous works, our work provides a more detailed set of annotations that include different clinical aspects of pain, as well as two other event types (treatment and anesthesia) important for studying outcomes. Similarly, there has not been any work on anesthesia-specific annotation and extraction.

Relating this work to a larger context, our pain, treatment, and anesthesia event annotations can be thought of as more specific reincarnations of the CLEF corpus and i2b2 event annotations (Roberts et al., 2008; Uzuner et al., 2011). For example, under the CLEF annotation schema, pain would fall under the condition entity, with the pain's location aligning to CLEF's locus/sub-location/locality schema. Drug, intervention, and negation for conditions are also elements we capture in our annotation schema. Under the i2b2/VA 2010 concepts, assertions, and relations challenge schema, pain would be considered a medical problem and pain treatments or anesthesia could be identified treatments. Our annotation of status' are related to assertion and relations between pain and treatment function similarly to their medical problem treatment relations. Pain and treatment annotation can also be compared to medication and adverse drug events, where instead the focus of events are on pain symptoms and treatment concepts (Uzuner et al., 2010; Karimi et al., 2015).

3 Corpus creation

We drew data from two sources (1) Stanford University's (SU) Clarity electronic medical record database, a component of the Epic Systems software, and (2) MTSamples.com, a online source of anonymized dictated notes. With approval of an institutional review board, we identified a cohort of surgical patients that underwent 5 procedures associated with high pain: distal radius fracture, hernia replacement, knee replacement, mastectomy, and thoracotomy. We focused on three note types: anesthesia, operative, and outpatient clinic visit notes. Anesthesia and operative notes were sampled from the day of surgery, whereas clinic notes were randomly sampled within 3 months prior and 1 year after the surgery. Because of the variation in clinic notes, we performed stratified random sampling per sub-note type and per surgery category.

From MTsamples, we isolated operative (surgery) and clinic visit notes. Clinic notes were considered those not grouped into specialized categories, e.g. surgery, autopsy, discharge. Frequencies by type are shown in Table 1.

Corpus	Anesthesia	Clinic	Operative
MTsamples	-	90	75
SU	90	90	75
TOTAL	90	180	150

Table 1: Breakdown of note types

4 Guideline Creation

Annotation guidelines were created iteratively with a medical general practitioner as well as a biomedical informatics scientist. The initial pain event schema was derived from existing literature (Fink, 2000) and cues from Stanford Health Care's pain collection practices. Schemas were designed and altered according to feedback from a surgical attendee and an anesthesiologist.

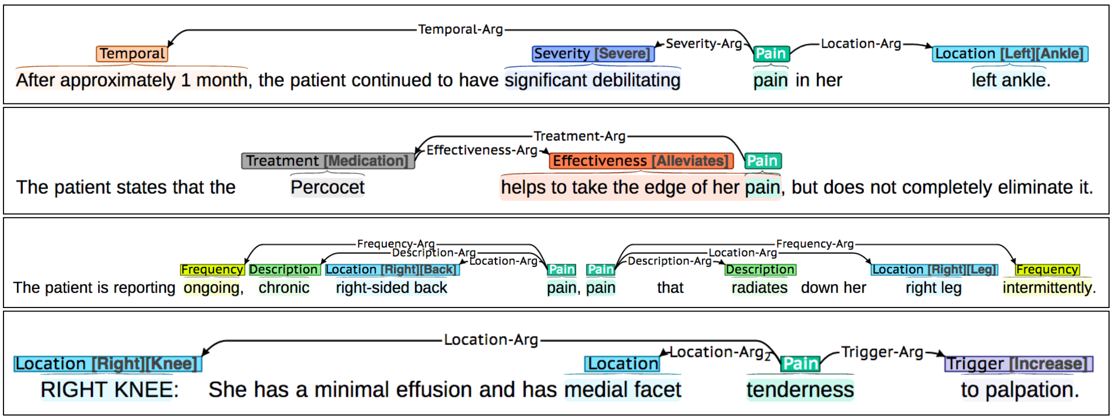
Our annotation focuses on three event types: pain, treatment, and anesthesia events. Below is a description of the entities (in some cases phrasal highlights) for each type of event. Those concepts marked with a * are event heads for which other entities may attach to.

Pain information:

Pain* - indication of pain including signs and symptoms that denote pain or diseases definitionally characterized as pain, e.g. *"myalgia"*, with attributes ***Goal:***{*binary*} and ***Status:***{*Current, Past, None, Unknown, Not Patient*}

Description - descriptive characteristics of the indicated pain, e.g. *"burning"*

Frequency - information regarding periodic oc-

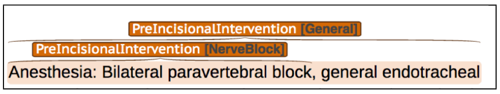

After approximately 1 month, the patient continued to have significant debilitating pain in her left ankle.

The patient states that the Percocet helps to take the edge of her pain, but does not completely eliminate it.

The patient is reporting ongoing, chronic right-sided back pain, pain that radiates down her right leg intermittently.

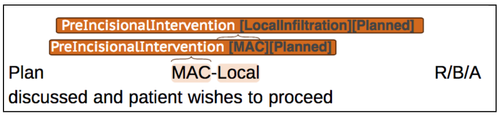
RIGHT KNEE: She has a minimal effusion and has medial facet tenderness to palpation.

Figure 1: Example pain and treatment events


Anesthesia: Bilateral paravertebral block, general endotracheal

PreIncisionalIntervention [General]
PreIncisionalIntervention [NerveBlock]

Figure 2: General and nerve block anesthesia text


Plan MAC-Local R/B/A discussed and patient wishes to proceed

PreIncisionalIntervention [LocalInfiltration][Planned]
PreIncisionalIntervention [MAC][Planned]

Figure 3: MAC and local anesthesia text

curence of the indicated pain, e.g. *"occasional"*
Location - location of pain, with attributes ***Laterality***:{*Bilateral, Left, Right, Unspecified*} and ***Type***:{*Abdomen, Ankle, Arm, Back, Back-lower, Back-upper, Breast, Buttocks, ChestArea, Ear, Elbow, Eye, Foot, Generalized, Groin, Hand, Head, Hip, Incisional, Jaw, Knee, Leg, Mouth, Neck, Nose, Pelvis, Shoulder, Throat, Wrist, Other*} (This attribute is useful for matching with structured data that pre-specify locations)
Severity - severity of pain, with attribute ***Severityattribute:*** {*0,1,..10, mild, moderate, severe* }
Temporal - demarkations of time points at which pain occurs, including time relative to events
Treatment - interventions used on patient (see next section for more information)
Trend - trend of pain with attribute ***TrendAttribute:*** {*Increasing, Decreasing, No change*}
Trigger - events that cause some change in pain, with attribute ***TriggerAttribute***:{*Increase,Decrease*}

Treatment information:
Effectiveness - Effectiveness of treatment with attributes ***EffectivenessAttribute:*** {*Alleviates, Worsens, No change*}
Treatment* - possible treatments for pain with attributes ***Type:***{*Acupuncture, Electrotherapy, Heat/cold therapy, Medication, No further action, Other, Physical Therapy, Steroid injection, Surgical procedure*} and ***Status:***{*Current, Past, None, Planned, Requested, Recommended, ConditionalRecommended, NotPatient*}
Temporal - demarkations of time points at which treatment occurs, including time relative to events

Anesthesia information:
Pre-incisional intervention* - anesthetic intervention that occurs prior to incision, with attributes ***Status***:{*Current, Past, None, Planned, Requested, Recommended, NotPatient*} and ***Type:***{*General, Regional-unspecified, Nerve block, Spinal block, Epidural, MAC (monitored anesthesia care), Local infiltration*}

Event heads, e.g. treatment, were always annotated whereas event arguments, e.g. effectiveness, were only annotated when an event head was present. Only pain medications defined in a curated list (or its synonyms) were annotated as treatment entities to avoid medical knowledge reliance. To avoid annotation fatigue, Status attributes were unmarked if Current.

5 Annotation

After development of an initial schema, a random sample of documents from each SU and MTSamples of anesthesia, operative, and clinical notes were drawn to measure inter-annotator agreement between a general practitioner and a biomedical informatics scientist. Pain and treatment events were annotated for clinical notes, whereas only pre-incisional intervention events were annotated

Field	Set1	Set2	Set1+2	Full
Description	1.00	0.250	0.625	36
Effectiveness	–	0.833	0.769	22
Frequency	0.889	0.909	0.900	36
Location	0.800	0.870	0.832	512
Pain	0.912	0.947	0.929	613
Severity	0.966	0.914	0.921	88
Temporal	0.500	0.698	0.628	200
Treatment	0.686	0.832	0.791	671
Trend	0.770	0.00	0.625	21
Trigger	0.884	0.851	0.839	128
ALL	0.797	0.858	0.831	2327

Table 2: IAA and counts for clinic note entities

Field	Set1	Set2	Set1+2	Full
EffectivenessAttribute	–	0.333	0.308	21
LateralityAttribute	0.758	0.804	0.774	101
LocationAttribute	0.737	0.716	0.700	457
Goal	–	0.920	0.911	16
Pain:StatusAttribute	0.756	0.885	0.822	201
SeverityAttribute	0.966	0.778	0.843	87
Treatment:Type	0.647	0.773	0.744	654
Treatment:StatusAttribute	0.595	0.569	0.597	499
TrendAttribute	0.769	0.00	0.625	21
TriggerAttribute	0.465	0.766	0.602	126
ALL	0.697	0.766	0.749	2183

Table 3: IAA and counts for clinic note attributes

for anesthesia and surgery notes.

An initial set (Set1) included 15 clinic and 15 operative notes from MTSamples; and 30 anesthesia, 15 clinic, and 15 operative notes from SU. Two rounds of revision and agreement were performed on this set. Changes or adjustments to annotation guidelines were made as necessary during annotator agreement cycles. Because clinic notes presented more complexity, we drew another 15 documents from MTSamples and 15 from SU resulting in a new subset (Set2). EffectivenessAttribute and Goal attributes were added from the second set onwards. Two rounds of revisions were performed on this set. Finally, the combined set was revised. The remaining corpus (60 anesthesia, 120 clinic, 120 operative notes) was evenly split and single-annotated by the two annotators. We used brat, a web-based software, for our annotation (Stenetorp et al., 2012).

Inter-annotator agreement (IAA) was evaluated using F1 measure, the harmonic mean of positive predictive value and sensitivity, for entities, relations, and attributes (Hripcsak and Rothschild, 2005). All reported measures are based on partial matches (text spans need only to overlap). For this, relations require that corresponding entity arguments overlap with accurate relation labels.

Field	Set1	Set2	Set1+2	Full
Description-Arg	0.667	0.250	0.533	38
Effectiveness-Arg	–	0.909	0.909	23
Frequency-Arg	0.923	0.769	0.846	37
Location-Arg	0.738	0.864	0.795	520
Severity-Arg	0.968	0.889	0.909	91
Temporal-Arg	0.449	0.738	0.620	221
Treatment-Arg	0.800	0.500	0.522	41
Trend-Arg	0.769	0.00	0.625	21
Trigger-Arg	0.883	0.773	0.800	131
ALL	0.744	0.797	0.760	1123

Table 4: IAA and counts for clinic note relations

Field	Set1	Set2	Set1+2	Full
Type	0.906	–	0.906	257
StatusAttribute	0.898	–	0.898	40
ALL	0.902	–	0.902	297

Table 5: IAA and counts for anesthesia note attributes

Field	Set1	Set2	Set1+2	Full
Type	0.935	–	0.935	237
StatusAttribute	0.860	–	0.860	5
ALL	0.897	–	0.897	242

Table 6: IAA counts for operative note attributes

6 Results

Tables 2-6 show final agreement levels for the separate sets of inter-annotator documents and then for the full inter-annotator corpus for the entities, attributes, and relation levels. We also report the frequencies of each field for the full corpus.

For clinic notes, 125 documents had at least one entity, with 19 ± 19 entities, 10 ± 11 relations per non-empty report. Table 7 shows the top 90% of unique co-occurring relation combinations attached to the same pain entity. Most pain entities appeared either without attached relations or with a Location-Arg. For treatment entities not attached to pain entities as an argument (632 entities), 74% had no attachments, 24% were attached to a Temporal-Arg alone, the rest had either an Effectiveness-Arg relation alone or both. Most relations existed within a close context, however a small number did appear at 2 or more sentences away. This included 10% of Trigger-Arg, 7% of Treatment-Arg, 2% of Severity-Arg, and 2% of Temporal-Arg relations. The remaining relations appeared on the same or one sentence away.

Identification of pain and treatment events for clinical notes was relatively challenging. Ten entities with their related attributes, as well as 8 relation types were involved. Moreover, clinical

Top co-occurring relations for same pain	Count	Fraction	Cum. Fract.
{Location-Arg}	285	0.465	0.465
{}	45	0.073	0.538
{Trigger-Arg}	35	0.057	0.595
{Location-Arg, Trigger-Arg}	28	0.046	0.641
{Location-Arg, Temporal-Argv}	26	0.042	0.684
{Severity-Arg}	22	0.036	0.719
{Location-Arg, Severity-Arg}	18	0.029	0.749
{Description-Arg, Location-Arg}	16	0.026	0.775
{Frequency-Arg, Location-Arg}	16	0.026	0.801
{Severity-Arg, Trigger-Arg}	12	0.020	0.821
{Location-Arg, Treatment-Arg}	9	0.015	0.835
{Temporal-Arg}	9	0.015	0.850
{Treatment-Arg}	8	0.013	0.863
{Trend-Arg}	8	0.013	0.876
{Location-Arg, Severity-Arg, Trigger-Arg}	7	0.011	0.887
{Effectiveness-Arg, Treatment-Arg}	5	0.008	0.896
{Location-Arg, Trend-Arg}	5	0.008	0.904

Table 7: Frequency of relation-combinations connecting to same pain entity

notes tend to contain unpredictable expressions, e.g. *"pain [...] waxing and waning"* or *"worse with hiking"*, and narrative information that spans over several sentences, the conclusion of which could communicate a resolved status. Thirteen out of 613 mentions of pain were attributed as past. Out of 126 marked TriggerAttributes, 114 were aggravating factors (Increase), with only 12 mentions of alleviating factors (Decrease). Interestingly, many severity attributes were qualitative descriptions with 22 for mild, 13 for moderate, and 23 for severe out of 87 total marked. For treatment types, of 654 identified treatment types, 428 were surgical procedures, 116 medication, 82 physical therapy, 12 steroid injection. The remaining had frequencies of 3-5 each.

Ideologically, there were nuances to annotating pain information. While the easiest references to pain were trivial, e.g. *pain*, some required referencing dictionaries, e.g. *myalgia*, or reading context, e.g. *discomfort*. Distinguishing between cause of and timing for pain was not always clear. For example, in *"pain is worse in the morning"* and *"pain [...] when running"*, both underlines could be considered as either Trigger or Temporal. Our final decision was to mark as a Trigger when believed to be causal of the pain rather than delineating chronology. Some pain attributes had multiple connotations. For example, *"chronic pain"*, defined as presence of pain for longer than 3 months, has both a duration and frequency context. We decided to assign *chronic* as a description attribute. Extent of decisions were specified in annotation guidelines. Finally, there are unavoidable limitations in text interpretation. For example, in *"patient is very tender to palpation"*, *very* may be normalized to moderate or severe based on annotator subjectivity. Furthermore, pain may be suggested but not explicitly stated, e.g. *"woman [...] with [...] debilitating abdominal wall hernias"* (most likely painful), and therefore not captured.

Anesthesia and operative note entity agreement was at 0.923 F1 and 0.934 F1. There was a total of 235 and 254 entities for anesthesia and operative notes. For anesthesia reports, 72 had at least one entity, with 4 ± 5 entities each; operative reports, 130 had at least one entity, with 2 ± 1 entities each. 15% of Pre-incisional intervention entities were marked as Planned for anesthesia reports; 1% for operative reports. Agreements for operative and anesthesia entities and attributes were high (Table 5 and 6). This is due to the focused nature of these domains. However, our annotation schema did not include implicit references, e.g. *"skin was anesthetized with 1% lidocaine solution"* where *lidocaine* is often used for local anesthesia.

To improve IAA, further annotation would benefit from pre-annotation of entities trained on this starting set. This would increase consistency and throughput. Additional annotation of a larger corpus would provide larger samples sizes to estimate task challenge for less populated classes.

7 Conclusions and Future Work

In this work, we present a rich annotation schema for pain and pain interventions, as well as an annotation categorization for anesthesia types. Although this work was developed in the surgical setting, the pain annotation schema presented here can be adapted for other settings. Future work includes building our extraction system and applying these data to assess important patient outcomes and health services research.

Annotation guidelines and the MTSamples portion of our corpus is available through our group's website (med.stanford.edu/boussard-lab.html).

Acknowledgments

This project was partially funded from the Veterans Affairs Big Data Scientist Training Enhancement Program. This project was supported by grant number R01HS024096 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality or the Department of Veterans Affairs.

References

Jaime L Baratta, Eric S Schwenk, and Eugene R Viscusi. 2014. Clinical consequences of inadequate pain relief: barriers to optimal pain management. *Plastic and reconstructive surgery* 134(4S-2):15S–21S.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research* 32(suppl 1):D267–D270.

Regina Fink. 2000. Pain assessment: the cornerstone to optimal pain management. In *Baylor University Medical Center. Proceedings*. Baylor University Medical Center, volume 13, page 236.

Norris H Heintzelman, Robert J Taylor, Lone Simonsen, Roger Lustig, Doug Anderko, Jennifer A Haythornthwaite, Lois C Childs, and George Steven Bova. 2013. Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. *Journal of the American Medical Informatics Association* 20(5):898–905.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association* 12(3):296–298.

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

Henrik Kehlet, Troels S Jensen, and Clifford J Woolf. 2006. Persistent postsurgical pain: risk factors and prevention. *The Lancet* 367(9522):1618–1625.

Karl A Lorenz, Cathy D Sherbourne, Lisa R Shugarman, Lisa V Rubenstein, Li Wen, Angela Cohen, Joy R Goebel, Emily Hagenmeier, Barbara Simon, Andy Lanto, et al. 2009. How reliable is pain as the fifth vital sign? *The Journal of the American Board of Family Medicine* 22(3):291–298.

Michael JE Neil and William A Macrae. 2009. Post surgical pain-the transition from acute to chronic pain. *Reviews in pain* 3(2):6–9.

Douglas Redd, Cynthia Brandt, Kathleen Akgun, Jinqiu Kuang, and Qing Zheng-Treitler. 2016. Improving pain assessment in medical intensive care unit through natural language processing. In *American Medical Informatics Association Annual Conference 2016*.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Andrea Setzer, and Ian Roberts. 2008. Semantic annotation of clinical text: The clef corpus. In *Proceedings of the LREC 2008 workshop on building and evaluating resources for biomedical text mining*. pages 19–26.

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*. Association for Computational Linguistics, pages 102–107.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association* 17(5):514–518.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* 18(5):552–556.

Jennifer F Waljee, Linda Li, Chad M Brummett, and Michael J Englesbe. 2017. Iatrogenic opioid dependence in the united states: Are surgeons the gatekeepers? *Annals of surgery* 265(4):728–730.

Paul F White and Henrik Kehlet. 2010. Improving postoperative pain managementwhat are the unresolved issues? *The Journal of the American Society of Anesthesiologists* 112(1):220–225.