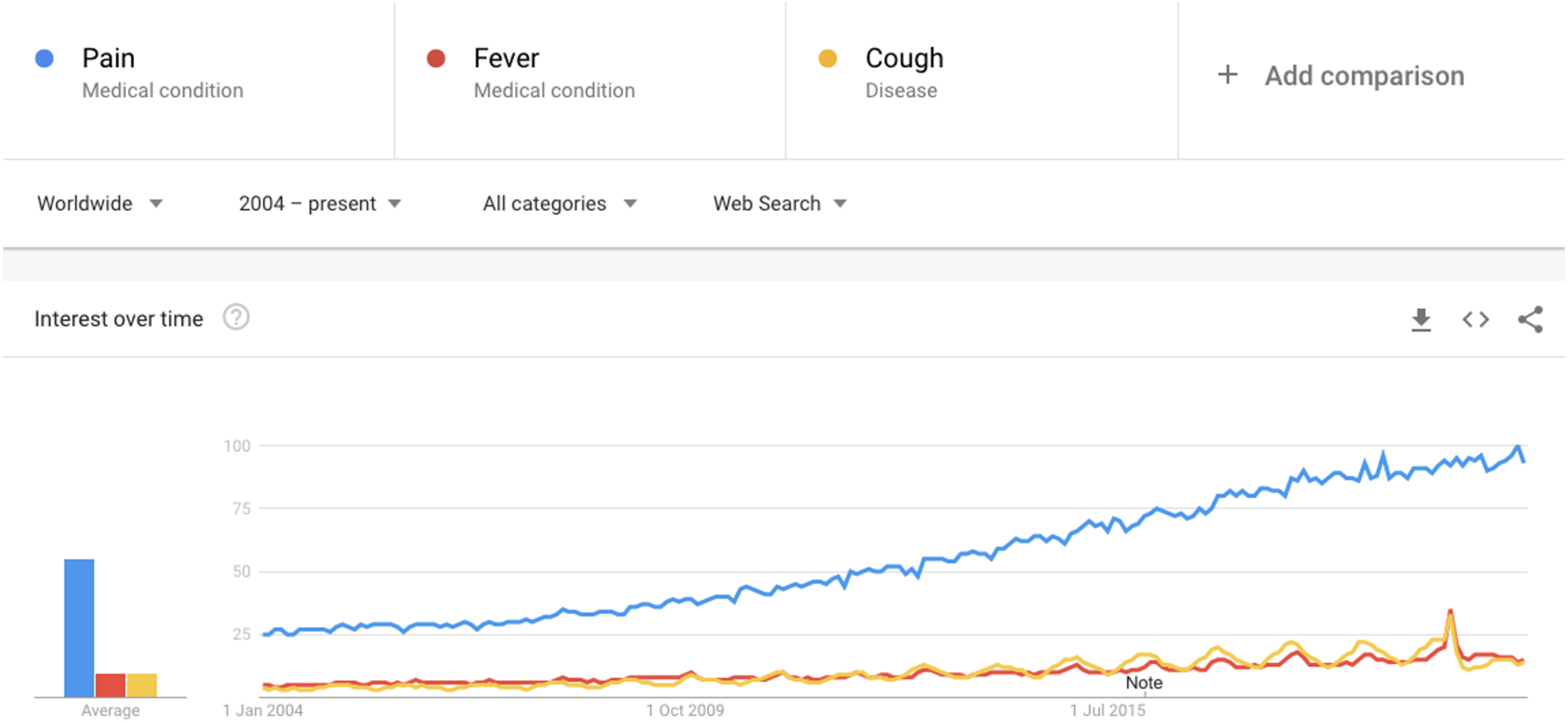
Development of a Lexicon for Pain

Jaya Chaturvedi
DRIVE-Health CDT PhD Student
King's College London

Pain

- Pain has been an active area of research, especially since the onset of the crisis of opioid use in the United States [1]
- Pain is known to have a strong relationship with emotions, which can lead to damaging consequences - worsened for people suffering with persistent pain [2]
- Apart from research, it has been of interest to the general population as well - google trends

Google trends - search term



● Pain
Medical condition

● Fever
Medical condition

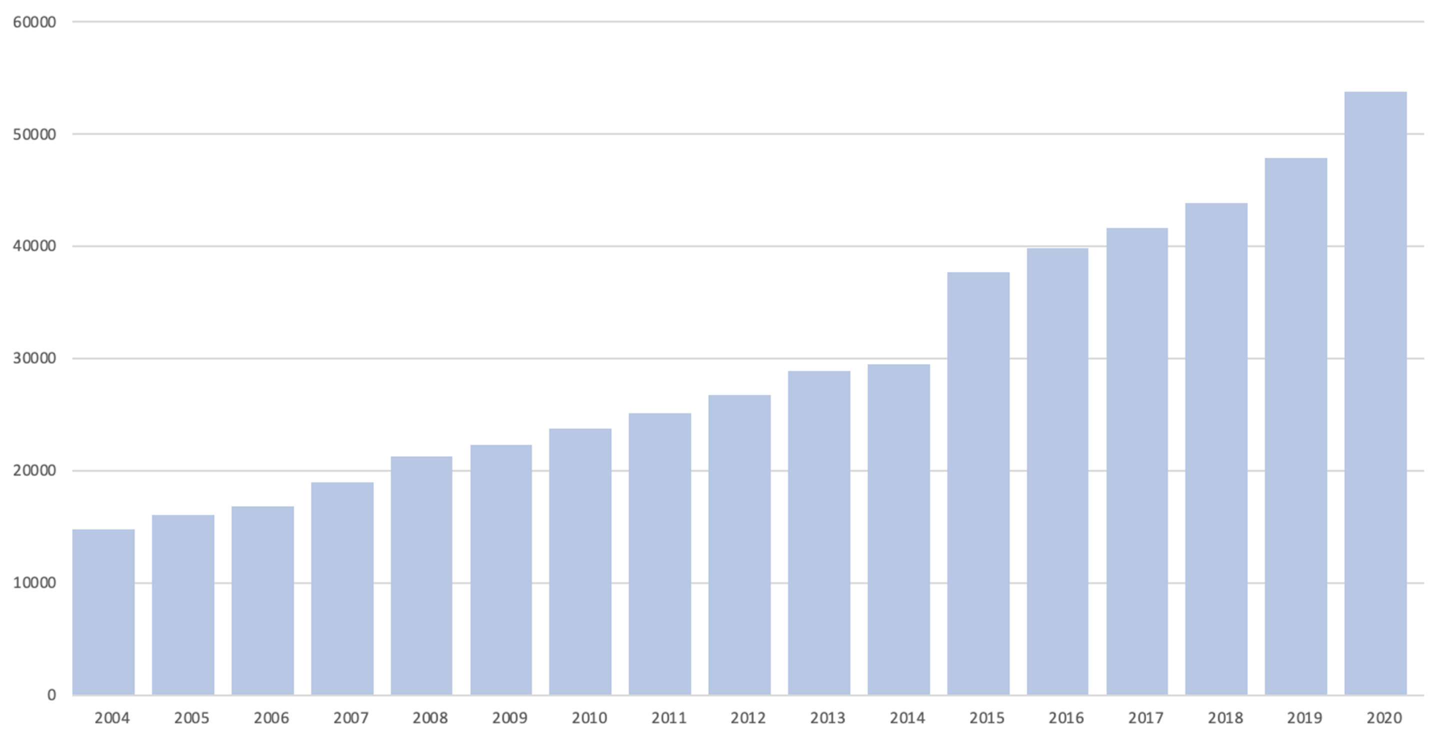
● Cough
Disease

＋ Add comparison

Worldwide ▾ 2004 – present ▾ All categories ▾ Web Search ▾

Interest over time

Publications for “pain” - Web of Science



Why am I interested in pain?

- The term “pain” presents a unique natural language processing (NLP) problem
  - subjective nature
  - ambiguous description
- Pain can refer to physical distress, or existential suffering, and sometimes metaphorical in phrases such as “for being a pain” [3]
- Within the biomedical context, it will most likely be the former two.

Purpose of this study - Develop an English lexicon of pain terms for use in NLP tasks

Exploration of Pain: Data Sources

1. Electronic Health Records
   a. CRIS
   b. MIMIC-III

2. Social Media
   a. Reddit
   b. Twitter

Electronic Health Records

CRIS database

- Anonymised version of EHR data from the South London and Maudsley NHS Foundation Trust (SLaM).
- Contains ~30 million event notes and correspondence letters, with an average of ~90 documents per patient
- Data extraction:
    - A SQL query was run on the attachments table of the CRIS database, and random 50 documents that contained the keyword ‘pain’ (both upper and lower case) were extracted.

Electronic Health Records

MIMIC-III

- Medical Information Mart for Intensive Care (MIMIC) is an openly available database which was developed by the Massachusetts Institute of Technology (MIT).
- De-identified healthcare data from the critical care units
- Discharge summaries and clinician progress notes average to about 3,900 words per document
- Data extraction:
  - A SQL query was run on the note-events table, and random 50 documents containing the keyword 'pain' (both upper and lower case) were extracted.

Social Media

Reddit

- A popular online community which supports unidentifiable accounts to allow users to post anonymously.
- Provides sub communities for people to discuss topics of shared interest (SubReddits)
- Chronic pain subreddit community - r/ChronicPain
- Data extraction:
  - Python package - PRAW
  - All posts from ChronicPain subreddit community (7,700 posts)
  - Random 50 used for analysis

Social Media

Twitter

- An online micro-blogging platform with an enormous number of users who post short (280 characters or less) messages, referred to as “tweets”, on topics of interest.
- Data extraction:
    - Python package - tweepy
    - All tweets containing words “chronic pain” (7,701 posts)
    - Random 50 used for analysis

A note on ethics and data access

- While data from Reddit and Twitter are publicly available, applicable ethical research protocols proposed by Benton et al. (2017) were followed in this study [4]
- No identifiable user data was used, and any sensitive direct quotes were paraphrased.
- Data from twitter is available through their API after approval of registration for access to this data, details of which can be found in their general guidelines and policies documentation [5]
- Data access information for CRIS [6] and MIMIC-III [7] are detailed on their respective websites.

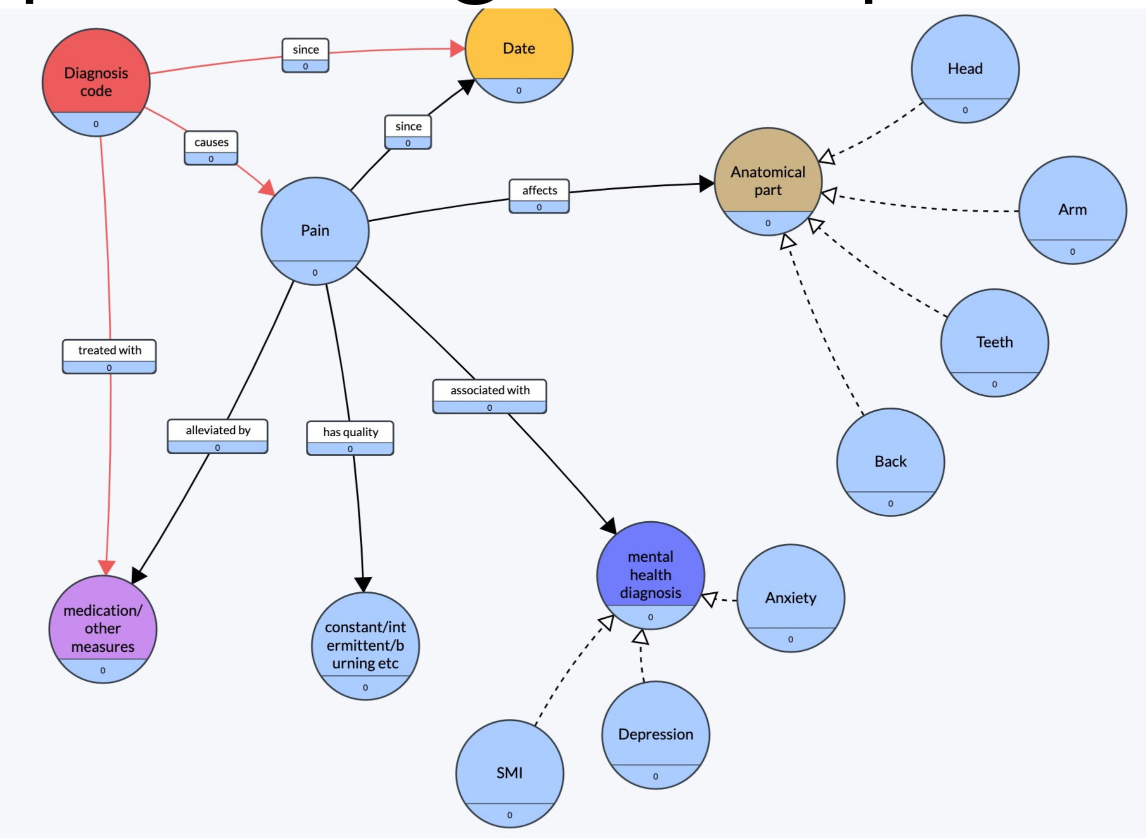
Collocates for "pain"

CRIS	MIMIC-III	Reddit	Twitter
chronic	control	pain	agony
back	acute	about	amazingly
clinic	chronic	anyone	achieved
physical	assessment	back	american
health	plan	anything	body

Common themes

Source	Type of pain	Feelings/experiences asso. with pain	Medications and other measures	Related to body parts
CRIS	...in constant pain.. ...ongoing pain... ...pain was quite severe..	...overwhelmed by chronic pain problems... ...fear of pain... ...pain causing distress.. ...struggles with chronic pain...	...drugs to numb the pain... ...pain relief medication not controlling the pain... ...side effects from pain relief medication...	...chronic back pain.. ...chest pain...
MIMIC-III	...severe pain... ...atypical pain...	-	...PO as needed for pain... ...taking narcotic pain medication... ...managed with IV pain medication...	...chronic back pain... ...chest pain... ...abdominal pain...
Reddit	Sharp pain.. Widespread pain..	...could be causing pain ...painful trips to kitchen ...am in the same painful position I was 3 months ago...	...helped my back pain...	...chronic neck pain.. ...back pain...
Twitter	-	...to live pain-free..	...muscle painbuster amazing discount!	...back pain

Conceptual diagram of pain



Building the Lexicon

Terms were generated from 3 different sources:

1. Literature
2. Ontologies
3. Embedding models

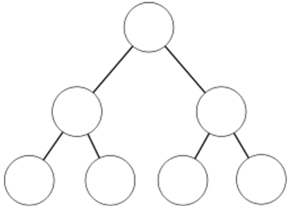
Literature

Pain-related words were harvested from three publications which made pain-related terms available in their supplemental materials:

1. A systematic review on application of NLP methods for symptom extraction from electronic patient-authored text (ePAT) [8].
2. A survey of biomedical literature based word embedding models [9].
3. A list of sign and symptom strings generated using NLP to meaningfully depict experiences of pain in patients [10].

These lists were cleaned by lowercasing all terms, and only keeping terms made up of 1 or 2 tokens.

Ontologies



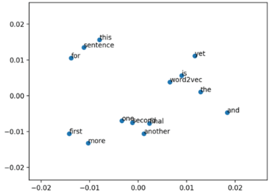
Synonyms for pain were incorporated from three biomedical ontologies:

1. The Unified Medical Language System (UMLS) [11],
2. Systematized NOmenclature of MEDicine Clinical Terms (SNOMED-CT) [12],
3. International statistical Classification of Diseases and related health problems: tenth revision (ICD-10) [13].

UMLS contains concepts from SNOMED-CT and ICD-10, in addition to a number of other vocabularies.

From each, we extracted terms of up to 2 tokens that either matched "pain*", were synonyms of pain, or described as child nodes of pain.

Embedding models



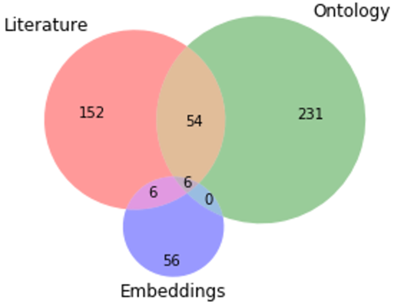
Embedding models [36, 37] using 8 different parameters and four different text sources were used to generate additional words similar to “pain”:

1. Two of the embedding models built using clinical text within MIMIC-II database.
2. Four embedding models built using clinical text within MIMIC-III (three using gensim implementation of word2vec and one using FastText.)
3. A model was built using word2vec over an SMI cohort from CRIS.
4. A publicly available model built on PubMed and PubMed Central (PMC) article texts

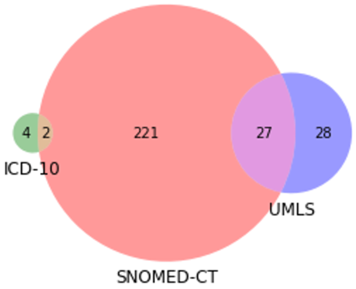
Only unigrams were included from the different models. Any duplicates were removed, and the remaining terms were added to the lexicon.

Lexicon coverage

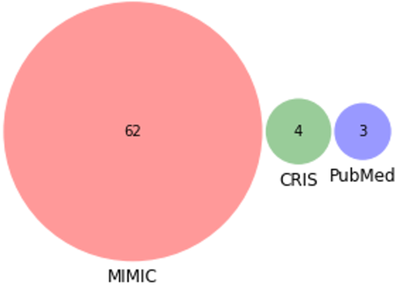
Comparison of different sources



Comparison of different ontologies



Comparison of different embeddings



Lexicon source	# of unique terms	Total # of terms
Literature	218	241
Ontologies	291	523
Embeddings	68	171

Validation of Lexicon

- The final list was validated by two clinicians, with the final size of the lexicon being 382 terms (30% unigrams, 68% bigrams, and 1% trigrams).
- Further validation was conducted by comparing the lexicon to another ontology which consisted of some pain terms - the Experimental Factor Ontology [14]:
    - 70% of the terms within the pain lexicon were also in the Experimental Factor Ontology, with all pain-related terms being covered (such as pain, ache, cramp)

Some patterns

1. **Anatomy**

   <anatomy term> <pain term> such as chest pain, ear ache, head discomfort

   <pain term><anatomy term> such as acute abdomen, aching muscles

1. **Quality of pain**

   <quality term><pain term> such as burning pain, chronic pain

   <pain term><quality term> such as pain burning, pain crushing

1. **Quality of pain and anatomy**

   <quality term><anatomy term><pain term> such as chronic back pain

Conclusion

- The final pain lexicon and the code to generate the embedding models is openly available on GitHub*
- The lexicon will also be added to other ontology collections such as BioPortal.
- This final lexicon will be used in downstream tasks such as building an NLP application to extract mentions of pain from clinical notes.

*https://github.com/jayachaturvedi/pain_lexicon

Acknowledgements

Supervisors:

Dr Angus Roberts

Dr Sumithra Velupillai

Clinicians:

Dr Rob Stewart

Dr Brendon Stubbs

Colleagues:

Dr Natalia Viani

Aurelie Mascio

Thank you!
Questions?

References

1 Howard R, Waljee J, Brummett C, Englesbe M, Lee J. Reduction in Opioid Prescribing Through Evidence-Based Prescribing Guidelines. JAMA Surgery. 2018 Mar 1;153(3):285–7.

2 Heintzelman NH, Taylor RJ, Simonsen L, Lustig R, Anderko D, Haythornthwaite JA, et al. 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. 2013 Sep 1;20(5):898–905.

3 Carlson LA, Hooten WM. Pain—Linguistics and Natural Language Processing. Mayo Clin Proc Innov Qual Outcomes. 2020 Apr 25;4(3):346–7.

4 Benton A, Coppersmith G, Dredze M. Ethical Research Protocols for Social Media Health Research. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing [Internet]. Valencia, Spain: Association for Computational Linguistics; 2017 [cited 2021 Mar 18]. p. 94–102. Available from: https://www.aclweb.org/anthology/W17-1612

5 https://help.twitter.com/en/rules-and-policies/twitter-api

6 https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/

7 https://mimic.physionet.org/

8 Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inform. 2019 May;125:37–46.

9 Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. A survey of word embeddings for clinical text. Journal of Biomedical Informatics: X. 2019 Dec 1;4:100057.

10 Heintzelman NH, Taylor RJ, Simonsen L, Lustig R, Anderko D, Haythornthwaite JA, et al. 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. 2013 Sep 1;20(5):898–905.

11 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;(32(Database issue)):D267-70.

12 Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp. 2001;662–6.

13 World Health Organization. ICD-10 : international statistical classification of diseases and related health problems : tenth revision. 2nd ed. World Health Organization; 2004. Spanish version, 1st edition published by PAHO as Publicación Científica 544.

14 Adamusiak T. Experimental Factor Ontology - pain [Internet]. OLS - Ontology Search. [cited 2021 Jun 2]. Available from: https://www.ebi.ac.uk/ols/ontologies/efo/terms?short_form=EFO_0003843