LENS: Learning Entities in Narratives of Skin Cancer

Daisy Monika Lal[1], **Paul Rayson**[1], **Christopher Peter**[2],
Ignatius Ezeani[1], **Mo El-Haj**[1], **Yufeng Liu**[3], **Yafei Zhu**[4]
[1]School of Computing and Communications, Lancaster University, UK.
[2] National Institute of Mental Health and Neuroscience, Bengaluru, India.
[3] Linguistics and English Language, Lancaster University, UK.
[4] Shanghai International Studies University, China.
Correspondence: d.m.lal@lancaster.ac.uk

Abstract

Learning entities from narratives of skin cancer (LENS) is an automatic entity recognition system built on colloquial writings from skin cancer-related Reddit forums. LENS encapsulates a comprehensive set of 24 labels that address clinical, demographic, and psychosocial aspects of skin cancer. Furthermore, we release LENS as a pip package[1], making it easy for developers to download and install, and also provide a web application[2] that allows users to get model predictions interactively, useful for researchers and individuals with minimal programming experience. Additionally, we publish the annotation guidelines[3] designed specifically for spontaneous skin cancer narratives, that can be implemented to better understand and address challenges when developing corpora or systems for similar diseases. The model achieves an overall entity-level F1 score of 0.561, with notable performance for entities such as "CANC_T" (0.747), "STG" (0.788), "POB" (0.714), "GENDER" (0.750), "A/G" (0.714), and "PPL" (0.703). Other entities with significant results include "TRT" (0.625), "MED" (0.606), "AGE" (0.646), "EMO" (0.619), and "MHD" (0.5). We believe that LENS can serve as an essential tool supporting the analysis of patient discussions leading to improvements in the design and development of modern smart healthcare technologies.

1 Introduction

Social-media channels have unlatched the way to modern healthcare by promoting patient engagement, professional communication, accessibility, education, and awareness (Sinclair et al., 2015; Aceto et al., 2018). These forums often assist patients and carers (family member, partner, or friend) when seeking advice, exploration and information gathering, sharing lived experiences, peer support, discussions, etc. (Naslund et al., 2020; Bruce et al., 2024). Together with online health communities (HealthUnlocked[4], AskaPatient[5], MedHelp [6], Health24 (Ji et al., 2023)), dedicated social health platforms (Reddit, Facebook, Twitter, YouTube) have revolutionised research and paved a way for modern medicine (Griffiths et al., 2012; Aase and Timimi, 2013; Moorhead et al., 2013; Aase and Timimi, 2013; Gupta et al., 2022). Particularly, patients with chronic illnesses such as cancer use social-media to seek out practical, social, and emotional support (Fox and Purcell, 2010; Patel et al., 2015; Foufi et al., 2019). Illness narratives, in which patients and carers recount their actual experiences, often in a chronological sequence that includes their past, present, and future, are potent forms of expression that benefit the listener on an emotional, social, and physical level (Charon, 2001, 2022).

Non-Medical Language	Medical Terminology
lump or bump	Tumor or Mass
unexplained weight loss	Cachexia
feeling tired all the time	Fatigue
shortness of breath	Dyspnea
bloating	Abdominal distension
multicolored spots	Pigmented Lesion
lumps in neck or groin	Lymphadenopathy
small, firm, red or pink bump	Papule

Table 1: Non-Medical Language vs Medical Terminology (Cancer-Related Symptoms)

At almost 40% of all cancer cases, skin cancer is one of the most common types of cancer globally

[1]*Installation instructions* available at https://github.com/dml2611/LENS.
[2]*Web App* available at https://lens-demo.streamlit.app/.
[3]*annotation_guidelines.pdf* available at https://github.com/dml2611/LENS/tree/main/annotation
[4]https://healthunlocked.com/
[5]https://www.askapatient.com/
[6]https://www.medhelp.org

Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations, pages 20–27
January 19–24, 2025.

Entity Type	Labels
Clinical	Cancer Type (CANC_T), Staging and grading (STG), Treatment (TRT), Part of Body (POB), Result (RES), Symptom (SYM), Investigation (INV), Medication (MED), Adverse Effect (ADV_EFF), Etiology (EGY), Tumor size and shape (SIZE), Number (NUM), Duration (DUR), Mental Health Diagnosis (MHD), Diagnosis of other diseases (DIAG), Organization (ORG), People or Cancer care team (PPL)
Demographic	Age (AGE), Gender (GENDER), Age/Gender (A/G), Geopolitical Entity (GPE)
Psychosocial	Emotion (EMO), Metaphorical Expression (MET), Other Expressions (EXP)

Table 2: Named entities in skin cancer narratives.

(Whiteman et al., 2016; Apalla et al., 2017; Bray et al., 2018; Urban et al., 2021). It is recorded as 17^{th} most common in males and 8^{th} most common in females, in Europe (Ferlay, 2004; WCRF, 2022). Biomedical named entity recognition (NER) can be used to extract key concepts (such as diseases, signs/symptoms, medications, treatment, side effects, gender, and age) from illness narratives (Kumar, 2020; Hao et al., 2021). It is a powerful technique that has gained attention in medical research communities for detecting named entities from clinical documents (Wen et al., 2021; Kocaman and Talby, 2022). Even though NER is one of the most valuable tools for information extraction, the lack of open-source cancer NER libraries is bottleneck for healthcare text analytics.

The major contributions of this work include:

1. **MELNER Corpus:** An annotated corpus including 24 named entity markers for skin cancer narratives. MELNER provides a detailed set of entities (see Table 2), including skin cancer-specific (cancer type, treatment, symptoms, etiology, etc.), demographic (age, gender, geographical location), and psychosocial (emotions, metaphors, and other expressions).

2. **LENS:** An NER system trained using the MELNER corpus, for automatic extraction of named entities from colloquial texts, such as narratives of patients and carers. LENS is made publicly available for distribution via PyPI and pip, making it easy for developers to download and install.

3. **Web Interface:** A website interface will be available (the streamlit hosted version is being developed), making it useful for researchers and individuals with minimal programming skills, for analysing and downloading the tagged entities.

4. **Label Mapping:** The LENS tags are further mapped to SNOMED-CT[7] and MedCAT[8] codes (see Table 7), useful for researchers and professionals in the medical domain.

5. **Annotation guidelines:** The annotation process, involving a wide range of labels, required a significant amount of effort, with the guide being amended frequently in agreement with the annotators and domain specialists. This evolved in a standardized system for labelling entities in skin cancer narratives. They may serve as relevant training material and promote collaboration with other researchers.

2 Motivation

Although multiple NER tools are being developed by researchers, their application to social-media data is still hindered by the usage of non-clinical expressions rather than domain-specific terminology (Denecke, 2014; He, 2019). Patients often resort to the use of figurative language for describing their symptoms (see Table 1). Additionally, errors and misspellings are common hallmarks of colloquial communication. The grammatical and lexical variability of health-related natural language on social-media poses a significant challenge for the NER task (Babaian and Xu, 2024). One of the significant drawbacks of the current biomedical NER systems is their narrow coverage, focussing only on clinical entities with fewer labels. Moreover, information such as social or demographic factors that are also connected to a patient's health is not considered by these systems (Raza et al., 2022). Analysing the psychosocial burden of the disease provides a more complete understanding of the patient experience and can help design more personalized treatment plans taking into account both medical and emotional aspects of care.

[7] https://www.snomed.org/what-is-snomed-ct
[8] https://github.com/CogStack/MedCAT

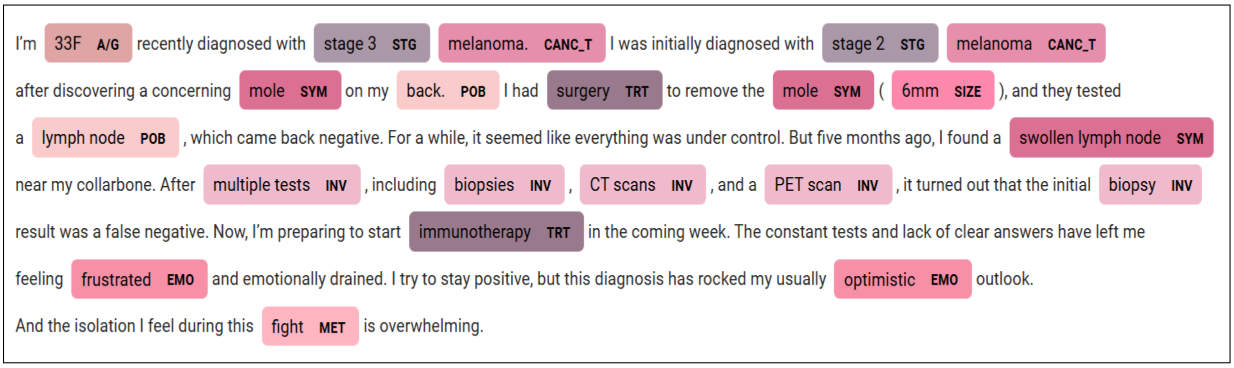

Figure 1: Extracting LENS entities from a narrative of skin cancer. For abbreviations see Table 2

3 Data Collection

Reddit supports a number of cancer-related forums that provide an anonymous environment for information sharing and discussions. As of January 2024, Reddit had 1.22 billion users globally, 73.1 million daily active users, and 36.7 million registered user accounts. There are roughly 100,000 active subreddits[9], with the cancer subreddit in particular having 58,380 subscribers[10].

We chose Reddit as the data source because, in addition to dedicated forums that reduce down manual searches for subjects of interest, it also allows for more structured text-based dialogue than other social media platforms. A finite number of 11 disease-specific subreddits relevant to skin cancer were empirically identified for analysis. These include, ‘r/cancer’, ‘r/skincancer’, ‘r/cancersurvivors’, ‘r/cancerfamilysupport’, ‘r/cancercaregivers’, ‘r/melanoma’, ‘r/melahomies’, ‘r/melanomasupport’, ‘r/melanomaquestions’, ‘r/dermatology’, and ‘r/dermatologyquestions.’

4 Annotations

We retrieved around 1000 posts, which were manually annotated by four interdisciplinary annotators from NLP, Linguistics, and the medical domain. The posts were as long as 800 tokens. The annotation task was carried out using the NER text annotator for SpaCy[11]. To avoid duplicates, which can lead to model overfitting, we assigned each annotator a mutually exclusive subset for annotation. This resulted in developing the final training set with 9435 identified named entities from 24 label categories (see Table 2).

[9]https://www.statista.com/
[10]https://subredditstats.com/r/cancer
[11]https://tecoholic.github.io/ner-annotator/

4.1 Annotation Guidelines

Before starting, the annotators were introduced to the annotation rules, annotation software, and the types of texts to be annotated. In addition to comprehensive guidelines, the annotators were advised to follow generic rules (see Table 3 for examples): (1) *Consistency:* Similar expressions must be annotated with the same label throughout the entire document. 2) *Precision:* Annotate effectively by selecting expressions that are highly related to the label and avoiding overly general terms. (3) *No overlap:* Avoid entity overlap unless stated otherwise (nested entities are not permissible). (4) *Punctuation Exclusion:* Avoid including punctuation marks in the named entity provided they are part of the expression. (5) *Complete Coverage:* Ensure that all designated categories are included in the narrative, where applicable. (6) *Reddit Slang:* Reddit frequently accepts abbreviations such as "33M" or "33/F", which represent a user’s age and gender. These shorthand expressions must be identified as unique entities. (7) *Medical Abbreviations:* Medical abbreviations like "WLE" for "Wide Local Excision" should be annotated consistently and not overlooked.

The annotation guidelines were periodically adjusted in response to issues encountered during the annotation process. The concerns were investigated with the guidance of domain experts, and rules were clarified with additional examples incorporated in the manual.

4.2 Inter-Annotator Agreement or IAA

The NER annotation is a sequence labeling task and bares numerous challenges, such as, (1) *Subjectivity:* Despite well-defined guidelines, annotations are susceptible to interpretation by the annotator and sometimes require a judgement call, leading to

Guideline	Correct ✓	Incorrect ✗
Consistency	"*chemotherapy*" → **TRT** (consistent across the document).	"*chemotherapy*" → **TRT** in one place, "*chemotherapy*" → **MED** elsewhere.
Precision	"*stage 2 melanoma*" → "*stage 2*" → **STG**, "*melanoma*" → **CANC_T**.	"*Stage 4 melanoma*" → **CANC_T** or **STG** without separating them.
No Overlap	"*swollen lymph node*" → **SYM** (without further annotation).	"*swollen lymph node*" → **SYM** for "*swollen lymph node*" and annotating "*lymph node*" separately as **POB** (overlapping).
Punctuation Exclusion	"*33 years old,*" → annotate "33 years old" as **AGE** (excluding the comma).	"*33 years old.*" → annotating "*33 years old,*" including the period as part of the entity.
Comprehensive Coverage	"*PET scan showed mets in the brain and lungs*" → "*PET scan*" → **INV**, "*mets*" → **INV**, "brain" → **POB**, "*lungs*" → **POB**.	Missing annotations for "*brain*" and "*lungs*" as **POB**.
Reddit Slang	"*I am 33M*" or "*I am 26/F*" → "*33M*" → **A/G**, and "*26/F*" → **A/G**	Annotating "*33M*" or "*26/F*" as **AGE** → **GENDER**.
Abbreviations	"*WLE*" → **TRT** or "*BRAF/MEK*" → **MED**	Not tagging the abbreviation.

Table 3: General guidelines for LENS annotation. For abbreviations see Table 2

conflicts, (2) *Uneven annotations*: the annotators frequently identify varying number of entities, (3) *Spanning*: annotators might not agree on the same spans for the same entity (see Table 4), (4) *Unannotated tokens*: Annotators may overlook tokens.

Due to the granularity of annotations, mismatches, and unannotated regions, the inter-annotator agreement using conventional metrics is quite challenging (Jiang et al., 2022; Dhrangadhariya et al., 2023). For instance, Cohen Kappa, a standard IAA metric, has been frequently pointed out by researchers as being inappropriate for assessing agreement on NER annotations (Hripcsak and Rothschild, 2005; Deleger et al., 2012; Karimi et al., 2015; Brandsen et al., 2020). We adopted an analogous approach to (Karimi et al., 2015) to compute the IAA. The agreement, $IAA(i, j)$, between two annotators, i and j, is defined as the average of absolute match ($abs(A_i, A_j)$) and fuzzy match ($fuz(A_i, A_j)$) as follows:

$$IAA(i,j) = \frac{abs(A_i, A_j) + fuz(A_i, A_j)}{max(A_i, A_j)} \quad (1)$$

Here, A_i represents the set of all annotations by the annotator i, A_j represents the set of all annotations by the annotator j, $max(A_i, A_j)$ denotes the maximum of A_i and A_j. $abs(A_i, A_j)$ counts exact matches, while $fuz(A_i, A_j)$ counts fuzzy matches or overlaps as specified in Table 5.

To evaluate the IAA, we employed a separate sample of five skin cancer narratives consisting of 210 sentences and 3814 words. Each annotator was requested to tag the narratives applying the same

('27 years', 'DUR')	('for 27 years', 'DUR')
('stage 4', 'STG')	('late stage 4', 'STG')
('frequent headaches', 'SYM')	('headaches', 'SYM')
('brain covering.', 'POB')	('brain', 'POB')
('cauterize the wound', 'TRT')	('cauterize', 'TRT')
('a few weeks ago', 'DUR')	('few weeks ago', 'DUR')
('was cut out', 'TRT')	('cut out', 'TRT')
('incredibly shocked', 'EMO')	('shocked', 'EMO')

Table 4: Example of fuzzy agreement where annotators identify different spans for the same label.

annotation guidelines. The number of labels identified by annotators ranges between 260 and 340. We observed that the annotators have an absolute match of 48-69%, whereas the fuzzy match adds another 6-14%. The IAA between annotators varies from 0.58 to 0.78. Because of the large number of labels and the size of the assessment set, this score indicates moderate to good agreement among the annotators.

5 Model Training

LENS was trained using the pre-trained SpaCY model with built-in pipelines, *ner* and *transformer (bert-base-cased)*. Training was conducted on an NVIDIA Tesla T4 with the Adam optimizer with a learning rate of 0.0001 and L2 weight decay of 0.01. Further configurations include a dropout rate of 0.01, dynamically adjusted batch sizes, and the F1 score was computed every 200 steps. The assessment metrics (see Table 6) illustrate an overall entity-level F1 score of 0.561, with notable

	Criteria	Examples
Absolute Agreement (Same span, same label)	(start1 = start2) and (end1 = end2) and (label1 = label2)	("chemotherapy", TRT) ("chemotherapy", TRT)
Fuzzy Agreement (Different spans, same label)	(start1 = start2) and (end1 ≠ end2) and (label1 = label2) or (start1 ≠ start2) and (end1 = end2) and (label = label2)	("pain", ADV_EFF) ("chronic pain", ADV_EFF)
Absolute Disagreement (Same span, different labels)	(start1 = start2) and (end1 = end2) and (label1 ≠ label2)	("chemotherapy", TRT) ("chemotherapy", MED)
Fuzzy Disagreement (Different spans, different labels)	(start1 = start2) and (end1 ≠ end2) and (label1 ≠ label2) or (start1 ≠ start2) and (end1 = end2) and (label ≠ label2)	("tumor", SYM) ("tumor spread", INV) or ("mole", SYM) ("removed the mole", TRT)

Table 5: Criteria for computing agreement on labels. Here, start, end, and label, denote the start index, end index, and label of the identified named entity by respective annotators.

Label	P	R	F1
CANC_T	**0.708**	**0.790**	**0.747**
STG	**0.712**	**0.881**	**0.788**
POB	**0.721**	**0.707**	**0.714**
TRT	0.650*	0.601*	0.625*
INV	0.525	0.473	0.497
SYM	0.487	0.361	0.415
A/G	**0.833**	0.625*	**0.714**
DUR	0.389	0.092	0.149
ORG	0.381	0.276	0.320
MED	0.649*	0.568	0.606*
AGE	**0.705**	0.596	0.646*
ADV_EFF	0.083	0.020	0.032
SIZE	0.273	0.188	0.222
RES	0.375	0.150	0.214
NUM	0.393	0.250	0.306
EMO	0.664*	0.580*	0.619*
MET	0.100	0.043	0.061
PPL	**0.732**	0.676	**0.703**
MHD	0.667*	0.400	0.500
GENDER	**0.750**	**0.750**	**0.750**
DIAG	0.375	0.158	0.222
EGY	0.250	0.250	0.250
GPE	0.286	0.333	0.308
EXP	0.000	0.000	0.000
Overall	0.618	0.514	0.561

Table 6: LENS model performance with precision (**P**), recall (**R**), and F1 Score (**F1**). Here, **bold** represents scores ≥ 0.70, * represents scores ≥ 0.60 and ≤ 0.69.

performance for entities such as "CANC_T" (F1: 0.747), "STG" (F1: 0.788), "POB" (F1: 0.714), "GENDER" (F1: 0.750), "A/G" (F1: 0.714), and "PPL" (F1: 0.703). Other entities with significant results include "TRT" (F1: 0.625), "MED" (F1: 0.606), "AGE" (F1: 0.646), "EMO" (F1: 0.619), and "MHD" (F1: 0.5).

6 SNOMED-CT and MedCAT Mappings

To facilitate interoperability, standardization, and clinical relevance with clinical records and databases, LENS entities are mapped to standardized systematised nomenclatures of medicine, SNOMED-CT (Spackman et al., 1997; Stearns et al., 2001; Cornet and de Keizer, 2008; Lee et al., 2014) and MedCAT (Fodeh et al., 2013; Kraljevic et al., 2021). This bridges the gap between patient-reported experiences and structured clinical information. The mappings were manually curated by identifying the most appropriate medical or conceptual label in each ontology, corresponding to the description and definition of LENS tags (see Table 7). The mappings[12] were cross-checked by a domain specialist to ensure that both mappings accurately represented the corresponding LENS tag. One drawback was that there were no comparable formal codes for metaphors and expressions.

7 Limitations

Several limitations impacted the efficacy of LENS: **(1) Limited Data:** While LENS performs well for some entities, tags like ADV_EFF, MET, and EXP exhibit low F1, highlighting the need for more

[12] Mappings *lens2snomedct.json* and *lens2medcat.json* available at https://github.com/4dpicture/LENS

Terms	LENS	SNOMED-CT	MedCAT
28/F	A/G	[Age, Gender finding]	[Temporal Concept, Organism Attribute]
stage 4	STG	Tumor staging	Clinical Attribute
melanoma	CANC_T	Malignant neoplastic disease	Disease or Syndrome
neck	POB	Body part	[Body Part, Organ, or Organ Component, Body Location]
swollen lymph node	SYM	Sign	Sign or Symptom
spread	INV	Investigations	[Diagnostic Procedure, Finding, Laboratory or Test Result]
frustrated	EMO	Emotions	Mental Process

Table 7: Examples of LENS tags mapped to SNOMED-CT and MedCAT.

training samples. However, medical annotation is costly and labor-intensive, and focusing on a single subcategory like skin cancer, further narrows the data pool due to limited forums. **(2) Variability in Expression:** Patients describe similar experiences in varied ways, making entity identification difficult. For example, "depressed" for an individual could mean clinical depression, while for another, it may simply mean a temporary low mood, leading to inconsistent tagging. **(3) Defining Annotation Guidelines:** Patients may express anything from diagnosis to remission using a multitude of expressions. This makes it challenging to decide what should or should not be included in the annotation guidelines, where it fits, and how to correctly label it. The process is tedious and seems never-ending, as new expressions are ever evolving, complicating the annotation process. **(4) Overlapping Entity Boundaries:** Entity boundaries can be ambiguous, making it difficult to determine the span and label, resulting in inconsistent annotations that confuse the model and degrade performance.

8 Conclusion

In this work, we introduce LENS, an open-source library for learning entities from narratives of skin cancer. LENS is an easy-to-use, production-ready model trained on spontaneous clinical narratives assembled from Reddit forums related to skin cancer. This research aims to assist healthcare professionals and oncology researchers, in the fast retrieval of information. In addition to medical concepts, LENS identifies demographic and psychosocial entities, often overlooked by traditional biomedical NER systems, narrowing down a critical gap in medical research and patient care. This work provides opportunities for research and collaboration, including (1) developing NER models to automatically extract clinical information from non-technical unstructured language, (2) enhancing patient-care by learning about psychosocial behaviour and mental health conditions cancer patients, (3) studying the impact of the disease in different age groups and genders using demographic labels, (4) mapping informal vocabulary to formal medical concepts such as SNOMED-CT and MedCAT, (5) using rigorous annotation guidelines for developing corpora and systems for similar diseases, (6) addressing challenges when dealing with personal narratives on social-media platforms, such as slang and medical abbreviations.

9 Ethics Statement

The large-scale analysis of personal narratives on open or closed online forums, particularly related to sensitive topics such as cancer, requires ethical approval, and we have been granted approval for secondary data analysis of previously analysed datasets. The research presented in this paper is part of a larger multilingual multinational research project, and each partner will apply it in their organization or country to replicate our analysis. The overall aim of the research is to improve the cancer patient journey and ensure personal preferences are understood and respected during treatment discussions with medical professionals, thereby supporting treatment and care choices at each stage of disease or treatment.

Acknowledgements

This publication presents research from the 4D PICTURE project,[13] which is a collaboration of research teams from Austria, Belgium, Denmark, Germany, the Netherlands, Slovenia, Spain, Sweden, and the UK. The research leading to these results has received funding from EU research and

[13]https://4dpicture.eu/

innovation programme HORIZON Europe 2021 under grant agreement 101057332 and by the Innovate UK Horizon Europe Guarantee Programme, UKRI Reference Number 10041120.

References

Lee Aase and Farris K Timimi. 2013. Health care social media: Engagement and health care in the digital era. *Clinical Obstetrics and Gynecology*, 56(3):471–476.

Giuseppe Aceto, Valerio Persico, and Antonio Pescapé. 2018. The role of information and communication technologies in healthcare: taxonomies, perspectives, and challenges. *Journal of Network and Computer Applications*, 107:125–154.

Zoe Apalla, Aimilios Lallas, Elena Sotiriou, Elizabeth Lazaridou, and Demetrios Ioannides. 2017. Epidemiological trends in skin cancer. *Dermatology practical & conceptual*, 7(2):1.

Tamara Babaian and Jennifer Xu. 2024. Entity recognition from colloquial text. *Decision Support Systems*, 179:114172.

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577.

Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. 2018. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424.

Emmanuel Bruce, Zhao Shurong, John Amoah, Sulemana Bankuoru Egala, and Francis Kofi Sobre Frimpong. 2024. Reassessing the impact of social media on healthcare delivery: insights from a less digitalized economy. *Cogent Public Health*, 11(1):2301127.

Rita Charon. 2001. Narrative medicine: a model for empathy, reflection, profession, and trust. *Jama*, 286(15):1897–1902.

Rita Charon. 2022. Narrative medicine. In *Palgrave Encyclopedia of the Health Humanities*, pages 1–11. Springer.

Ronald Cornet and Nicolette de Keizer. 2008. Forty years of snomed: a literature review. *BMC medical informatics and decision making*, 8:1–6.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, Imre Solti, et al. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, volume 2012, page 144. American Medical Informatics Association.

Kerstin Denecke. 2014. Extracting medical concepts from medical social media with clinical nlp tools: a qualitative study. In *Proceedings of the fourth workshop on building and evaluation resources for health and biomedical text processing*, pages 54–60.

Anjani Dhrangadhariya, Roger Hilfiker, Karl Martin Sattelmayer, Nona Naderi, Katia Giacomino, Rahel Caliesch, Julian Higgins, Stéphane Marchand-Maillet, and Henning Müller. 2023. Robuster: A corpus annotated with risk of bias text spans in randomized controlled trials. *Journal of Medical Internet Research*.

J Ferlay. 2004. Globocan 2002: cancer incidence, mortality and prevalence worldwide. *IARC cancer base*.

Samah Jamal Fodeh, Maryan Zirkle, Dezon Finch, Ruth Reeves, Joseph Erdos, and Cynthia Brandt. 2013. Medcat: A framework for high level conceptualization of medical notes. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 274–280. IEEE.

Vasiliki Foufi, Tatsawan Timakum, Christophe Gaudet-Blavignac, Christian Lovis, and Min Song. 2019. Mining of textual health information from reddit: Analysis of chronic diseases with extracted entities and their relations. *Journal of medical Internet research*, 21(6):e12876.

Susannah Fox and Kristen Purcell. 2010. *Chronic disease and the internet*. Pew Internet & American Life Project Washington, DC.

Frances Griffiths, Jonathan Cave, Felicity Boardman, Justin Ren, Teresa Pawlikowska, Robin Ball, Aileen Clarke, and Alan Cohen. 2012. Social networks–the future for health care delivery. *Social science & medicine*, 75(12):2233–2241.

Poonam Gupta, Asma Khan, and Amit Kumar. 2022. Social media use by patients in health care: a scoping review. *International Journal of Healthcare Management*, 15(2):121–131.

Tianyong Hao, Zhengxing Huang, Likeng Liang, Heng Weng, Buzhou Tang, et al. 2021. Health natural language processing: methodology development and applications. *JMIR medical informatics*, 9(10):e23898.

Zhe He. 2019. Understanding and bridging the language and terminology gap between health professionals and consumers using social media. *Social Web and Health Research: Benefits, Limitations, and Best Practices*, pages 103–121.

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.

Geonwoo Ji, Constantino Msigwa, Denis Bernard, Geon Lee, Jiyoung Woo, and Jaeseok Yun. 2023. Health24:

Health-related data collection from wearable and mobile devices in everyday lives. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 336–337.

Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. Annotating the tweebank corpus on named entity recognition and building nlp models for social media analysis. *arXiv preprint arXiv:2201.07281*.

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.

Veysel Kocaman and David Talby. 2022. Accurate clinical and biomedical named entity recognition at scale. *Software Impacts*, 13:100373.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, et al. 2021. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083.

A Kumar. 2020. Disambiguation model for bio-medical named entity recognition. *Deep Learning Techniques for Biomedical and Health Informatics*, pages 41–55.

Dennis Lee, Nicolette de Keizer, Francis Lau, and Ronald Cornet. 2014. Literature review of snomed ct use. *Journal of the American Medical Informatics Association*, 21(e1):e11–e19.

S Anne Moorhead, Diane E Hazlett, Laura Harrison, Jennifer K Carroll, Anthea Irwin, and Ciska Hoving. 2013. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of medical Internet research*, 15(4):e1933.

John A Naslund, Ameya Bondre, John Torous, and Kelly A Aschbrenner. 2020. Social media and mental health: benefits, risks, and opportunities for research and practice. *Journal of technology in behavioral science*, 5:245–257.

Rajesh Patel, Tammy Chang, S Ryan Greysen, and Vineet Chopra. 2015. Social media use in chronic disease: a systematic review and novel taxonomy. *The American journal of medicine*, 128(12):1335–1350.

Shaina Raza, Deepak John Reji, Femi Shajan, and Syed Raza Bashir. 2022. Large-scale application of named entity recognition to biomedicine and epidemiology. *PLOS Digital Health*, 1(12):e0000152.

Wendy Sinclair, Moira McLoughlin, and Tony Warne. 2015. To twitter to woo: Harnessing the power of social media (some) in nurse education to enhance the student's experience. *Nurse education in practice*, 15(6):507–511.

Kent A Spackman, Keith E Campbell, and Roger A Côté. 1997. Snomed rt: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association.

Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. 2001. Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association.

Katelyn Urban, Sino Mehrmal, Prabhdeep Uppal, Rachel L Giesey, and Gregory R Delost. 2021. The global burden of skin cancer: A longitudinal analysis from the global burden of disease study, 1990–2017. *JAAD international*, 2:98–108.

WCRF. 2022. Skin cancer statistics. *https://www.wcrf.org/cancer-trends/skin-cancer-statistics/.*

Chaojie Wen, Tao Chen, Xudong Jia, and Jiang Zhu. 2021. Medical named entity recognition from unlabelled medical records based on pre-trained language models and domain dictionary. *Data Intelligence*, 3(3):402–417.

David C Whiteman, Adele C Green, and Catherine M Olsen. 2016. The growing burden of invasive melanoma: projections of incidence rates and numbers of new cases in six susceptible populations through 2031. *Journal of Investigative Dermatology*, 136(6):1161–1171.