9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

The Online Medical Taxonomy

Keywords: NLP, social media analysis, health discussions, online medical taxonomy

Extended Abstract

Introduction The motivations to mine online discussions for tracking the outbreaks and evolution of infectious diseases and chronic conditions have been strengthened by the COVID-19 pandemic. To broaden the set of diseases that state-of-the-art algorithms can detect from the text, we developed a deep learning tool for Natural Language Processing that extracts mentions of virtually any medical condition or symptom from unstructured social media text.

Social media offer a cheap and real-time alternative source of data to paint "a fuller picture" of people's own health experiences. However, as suggested by [LKKV14], we should abandon the *"assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis,"* and we should develop methods that blend this data with official data (i.e., health agencies collect prevalence statistics, and periodic health surveys). Another issue that needs to be addressed is "*algorithm dynamics*," i.e., whether *"the instrumentation is actually capturing the theoretical construct of interest"* [LKKV14]. The final issue speaks to the *need for broad health measures* – many social media studies focus only on narrow yet important outcomes – health, however, encompasses a much broader range of aspects.

Our work partly tackles these challenges by: *1)* automatically deriving health taxonomy from social media discussions, *2)* proposing health metrics for a variety of uncovered conditions in this taxonomy that can be blended with official data; *3)* computing each condition's metric based on the limited set of symptoms related to that condition without over-fitting on unrelated terms; and *4)* proposing broader health metrics, making it possible to examine multiple conditions simultaneously.

Data and Methods Reddit is a public discussion website with communities (subreddits) dedicated to a broad range of themes, including health and well-being topics (e.g., /Depression, r/HealthyFood, r/Fitness). After applying MedDL [vMLQB20] tool to 141M Reddit posts, we built a co-occurrence network in which nodes are the extracted medical mentions from Reddit, undirected edges connect those mentioned in the same message, and edge weights are equal to the number of co-mentions. We analyzed the cluster structure of the resulting co-occurrence network of conditions using Infomap [RB08], and found that the clusters correspond to well-defined medical conditions.

Results By leveraging the hierarchical nature of these clusters, we created the first taxonomy of medical conditions automatically derived from online discussions (Table 1).

Validity of the Health Taxonomy: To assess the breadth of our taxonomy and to test whether its categories cover well-studied medical conditions, we compared it to the official International Classification of Diseases(ICD-11) of the World Health Organization (WHO), which contains 22 top-level disease categories, further split into sub-categories at multiple hierarchical levels. In this classification, diseases are organized mainly based on the body parts they concern. We matched our level-1 categories to the top-level ICD categories by simply searching the level-1 category on ICD. Out of our 34 level-1 categories, as many as 31 found

a match (Table 1). Those that did not span multiple ICD categories; for example, our *elderly* category contains conditions frequent among elderly people; yet, since these conditions affect different parts of the body, they are listed across multiple ICD categories. On the other hand, out of the 22 ICD categories, 20 are present in our taxonomy, making it the most extensive social data-driven categorization of medical conditions.

Validity of Health Scores: To best match our level-1 categories, from CDC, we gathered state-level prevalence statistics for arthritis, asthma, and self-reported 'mentally unhealthy days' and 'poor health', compiled between 2016 and 2017. From SAMHSA, we collected statistics on the prevalence of: mental illnesses, abuse of different substances (e.g., heroin), conditions linked to metabolic syndrome (e.g., diabetes prevalence), and Sexually Transmitted Diseases (STDs). We tested that the prevalence of the 18 specific health conditions i measured by official statistics negatively correlates with the corresponding health score H_i^l. The correlation results varied between -.19 and -.45, and suggest that our indices can capture real-world prevalence at varying degrees, and that the knowledge of the structure of the co-occurrence network is useful. The correlations between our health indices and official prevalences are not high. This gap can be explained by not all patients discussing their conditions online, certain states' populations being more tech-savvy, and some conditions being more likely to be discussed online than the others. This, however, opens up new avenues for future research such as understanding which conditions are over/under-represented on certain platforms, and that is key in designing an integration of our health indices with official health surveys and other health surveillance systems. Overall taxonomy structure can be compared to the human symptoms-disease network [ZMBS14] derived from scientific publications, to understand which novel links or clusters emerge from the social discussions.

Our methodology opens the path to systematically study the perceived health impact of diseases in large populations while broadening the opportunity to conduct digital health surveillance on medical conditions that have so far been overlooked.

References

[LKKV14] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(6176):1203–1205, 2014.

[RB08] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences (PNAS)*, 105(4):1118–1123, 2008.

[vMLQB20] Sanja Šćepanović, Enrique Martín-López, Daniele Quercia, and Khan Baykaner. Extracting medical entities from social media. In *Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL)*, pages 170–181, 2020.

[ZMBS14] XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. Human symptoms–disease network. *Nature Communications*, 5:4212, 2014.

(A) Level-1	*Level-2*	*Example words*
	Mental	
mental [06]	isolation, autism, adhd, bipolar, psychosis, severe illness, anhedonia, stress, tic, paranoia, anguish, dyslexia, depression, personality disorder	wobbly feel, dread, hypomania, autism, suicidal thought
anxiety [06]	anxiety	anxiety, anxious, panic attack
personality [06]	bpd, dysphoria, narcissistic, antisocial, schizotypal	lack of empathy, sociopathic, manipulative behaviour, abusive behaviour
	Behaviour	
breathing [12]	asthma, fatigue, chest, heart, breathing, active breathing control, inflammations	trouble breathing, severe chest pain
vomit [21]	vomiting, emetophobia, bugs, pain, gagging	terrible fever, phobic, disgust
STDs [01]	stds, yeast, pregnancy, pain	hiv, syphilis, viral load, testicular ache
obesity [05]	eating disorders, hunger, weight loss	obese, overweight, excessive fat, overeating
addiction [06]	drugs, porn, alcohol, symptoms	drinking problem, opiates, strong urge, abscess
sleep [07]	hallucinations, traumas, nightmares, apnea, narcolepsis, insomnia, sleepwalking	ptsd, flashback, apnea, snore, wake up every hour
	Body parts	
skin [14]	acne, redness, wrinkles, hyperpigmentation, scalp, aging, dryness, eczema, allergies, bites, herpes, food allergies, soreness, bumps, psoriasis, irritation, scab	pimple, whitehead, flaky, dark spot, ingrown hair, mango allergy
ear [10]	tinnitus, dementia, vertigo, vibrations, congestion, noise	ringing in my ear, dizzy, blowing nose constantly
eye [09]	vision distortion, blurry vision, gallstone, high pressure, eye alignment, blindness, glaucoma, sweating, light sensitivity, strain, hypertension, aneurysm, migraine	eye pressure, spatially aware, nearsighted
heart [11]	palpitations, irregular, tachycardia	irregular heartbeat, poor concentration
spine [08]	multiple sclerosis, neurogenerative, hernia	tingling, lesion, difficult to lay
back [15]	pain, sciatica, arthritis, lower, stiffness, dullness	hip pain, muscle stiffness, unable to sit up straight
reproductive [16,17]	stones, infections, clots, lupus, bladder	shave, pain with sex, extremely bloated
Level-1	***Level-2***	***Example words***
	Conditions	
cancer [02]	cancer, gout, skin, lymphoma, lumps, genitals, digestive, lymphnodes, bones	discolored skin, swollen lymphnode, back ache, terminally ill, gnarly bruise
infective [01]	sepsi, heart, fever, overdose, penumonia, mosquito-borne, measles, blood, confusion	highly viral, dark mucus, sweating and cough, with knuckle, blood clot
influenza [01]	viral, flu, yellow fever	increased temperature, loss of appetite
diabetes [05]	diabetes, cataract, metabolic syndrome, vision, brain	nebula, brain fog, low blood sugar, lost pigment
parkinson [08]	parkinson	tremor, jittering
injuries [22]	body, broken, nagging, traumas, head, disorientation	concussion, skull fracture, opiates
parasites [01]	lyme, fungi, fatigue, sleepiness	debilitating fatigue, fungal infection, dark spot
epilepsy [08]	seizure, spine, paralysis	spaced out feel, numb, muscle twitch
	Demographics	
female [16]	pcos, hair loss, vagina, cyst, endometriosis, pelvic, ovaries, spasm, menopause	hot flash, irregular period, swollen, cyst
infants [18]	reflux, ppd, breast, teeth	spitting up, clogged duct, nipple damage, screaming, mentally drained
elderly [-]	arthritis, prostate, hernia	urinary issue, cystitis, struggling to walk
pregnancy [18, 19]	birth complications, contractions, pms, shake/ache	regular contractions, bleeding, painful cramp
developmental [20]	birth defects, down syndrome, genetic, edema, preeclampsia, cystic fibrosis	absent nasal bone, unable to digest
	Systems	
nervous [08]	migrain, stroke, nerve pain, hemicrania, neck pain, persisting hallucinations	vessel occlusion, allodynia, cephalgia
respiratory [12]	cough, ear infection, sinus, sneezing, head, bronchitis, dryness, throat	sniffle, lingering cough, runny nose, tight airways, sore throat, abdominal discomfort
autonomic [-]	hypermobility, fibromyalgia, dysautonomia, erythema, patellofemoral, vasovagal	hard skin, spasm, fainting, arrhythmia
digestive [13]	bloating, hemorroid, irritation, bowel inflammation, celiac, constipation, diarrhea, gastritis	flare, trouble pooping, anal fissure
tyroid [05]	hypothyroidsm, burning mouth, hashimoto, infections, gastroparesis	lose my hair, growling stomach, swollen tyroid

(C) ICD-11 categories

[01] Certain infectious or parasitic diseases; **[02]** Neoplasms; **[03]** Diseases of the blood or blood-forming organs; **[04]** Diseases of the immune system; **[05]** Endocrine, nutritional or metabolic diseases; **[06]** Mental, behavioural or neurodevelopmental disorders; **[07]** Sleep-wake disorders; **[08]** Diseases of the nervous system; **[09]** Diseases of the visual system; **[10]** Diseases of the ear or mastoid process; **[11]** Diseases of the circulatory system; **[12]** Diseases of the respiratory system; **[13]** Diseases of the digestive system; **[14]** Diseases of the skin; **[15]** Diseases of the musculoskeletal system or connective tissue; **[16]** Diseases of the genitourinary system; **[17]** Conditions related to sexual health; **[18]** Pregnancy, childbirth or the puerperium; **[19]** Certain conditions originating in the perinatal period; **[20]** Developmental anomalies; **[21]** Symptoms, signs or clinical findings, not elsewhere classified; **[22]** Injury, poisoning or certain other consequences of external causes

Table 1: (*A*) The taxonomy of medical conditions extracted from Reddit, arranged in two levels, with some examples of individual entities. The numbers next to the level-1 category names correspond to the matched top-level ICD-11 categories. (*B*) The list of all top-level categories from ICD-11.