CREATING A HEALTH TAXONOMY WITH SOCIAL MEDIA

Anonymous authors
Paper under double-blind review

ABSTRACT

The motivation to mine online discussions for tracking the outbreak and evolution of diseases and chronic conditions has been strengthened by the Covid-19 pandemic. To broaden the set of diseases that state-of-the-art algorithms can detect from 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. After applying the tool to 141M Reddit posts, we analyzed the cluster structure of the resulting co-occurrence network of conditions, and found that the clusters correspond to well-defined medical conditions. By leveraging the hierarchical nature of these clusters, we created the first taxonomy of medical conditions automatically derived from online discussions. Based on the mentions of our taxonomy's sub-categories on geo-referenced Reddit posts, we computed disease-specific health scores at the level of U.S. states, and found that they correlated strongly with official statistics for 18 conditions. 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.

1 ONLINE HEALTH DISCUSSIONS

To monitor physical and mental health interventions, National Health agencies collect prevalence data for a broad range of diseases. However, such measurements do not paint "a full picture" of people's own health experiences. What are the patients' concerns? Which symptoms do they experience and how those symptoms evolve? To get a richer picture, some countries conduct periodic health surveys. However, these surveys are *1)* temporally coarse-grained, *2)* costly, and *3)* suffer from recall and reporting biases (Gkotsis et al., 2017). Social media offer a cheap and real-time alternative source of data. However, as suggested by Lazer et al. (2014), 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. Another issue that needs to be addressed is "*algorithm dynamics*," i.e., whether *"the instrumentation is actually capturing the theoretical construct of interest"* (Lazer et al., 2014). 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. Despite all symptoms and diseases are connected by a network of complex relationships (Zhou et al., 2014), most health-related social media studies have so far focused on individual diseases. This is partly due to the historical difficulty in developing text mining models that generalize across multiple health domains.

Our work partly tackles these three issues by: *1)* automatically deriving health taxonomy from social media discussions, *2)* proposing health metrics for a variety of uncovered conditions 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.

2 DATA AND METHODS

Social Media Dataset. 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). Users can post *submissions* to any subreddit, and add *comments* to submissions or to existing comments. We gathered all the posts made during the year 2017, for a total of 96M posts from 14M users (Baumgartner et al., 2020). To match Reddit discussions with official

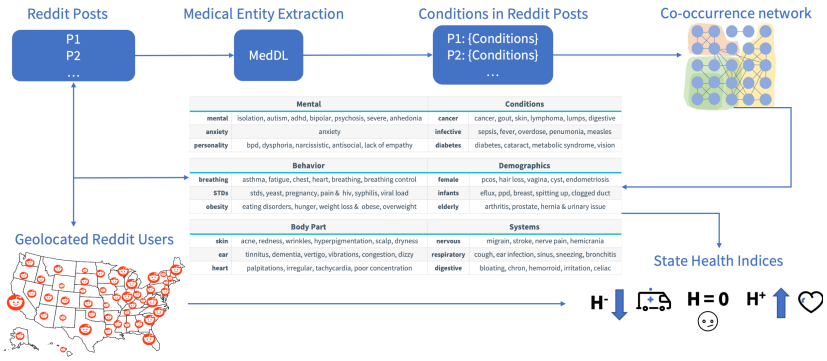

Figure 1: **Methodology everview**: Starting from Reddit posts of geolocated users, we extracted medical entities, then built the co-occurrence network, from which the heath taxonomy is created, which enables calculating the health indices across states. A concise versions of our taxonomy is shown in the middle. For the full version, see Appendix Table 2

health data, we focused on users we could locate at the level of the U.S. states. Reddit does not provide explicit user location, yet it is possible to get reliable location estimates with simple heuristics. Following previous work (Balsamo et al., 2019), we obtained a list of 484K users who are likely to be located in one of the 50 states (Pearson correlation $r = .95$ and $p < e^{-23}$ between the number of located users and state population). In 2017, these users authored 7M submissions and 134M comments.

Extracting Medical Entities from Social Media. To extract health mentions from social media text, we developed an NLP deep learning tool called MedDL (Šćepanović et al., 2020). The tool is based on deep recurrent neural networks (Huang et al., 2015) and contextual embeddings (Liu et al., 2019), and it was trained to extract mentions of any type of symptom from free-form text. The method achieved the strict F1 score for named entity recognition of symptoms ranging from .71 to .78 on three benchmark datasets, and outperformed previous baselines. Among the pre-trained MedDL models,[1] there is one for Reddit, which we used.

Health Taxonomy from Social Media. 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. This co-occurrence network captures the semantic relatedness of those mentions: those that appeared often together are likely to describe the same condition and form a densely-connected cluster of nodes. To find these semantically cohesive clusters, we tested several *community detection* algorithms, and in the end selected Infomap (Rosvall & Bergstrom, 2008). The density of co-occurrence networks is typically high, affecting the performance of these algorithms. Hence, we sparsified the network beforehand Coscia & Neffke (2017) and focused our analysis on the giant connected component with $411k$ nodes.

Upon application on the Reddit co-occurrence network, Infomap uncovered a health taxonomy: Figure 1 shows a concise and Appendix Table 2 the full version of it. There are 34 level-1 categories of medical conditions and 241 level-2 categories. These categories cover a wide range of medical conditions. We manually named the level-2 categories after inspecting the 50 most frequently used words they contained; we then manually named the level-1 categories based on the level-2 categories they contained. Finally, we grouped the level-1 categories into six main themes (greyed rows in Appendix Table 2). That is, symptoms associated with: mental health (the subgraph for this category is presented in Figure 2); individual body parts (e.g., eyes); systems of the human body (e.g., digestive system); specific demographics (e.g., women, elderly); various behaviours (e.g., eating); or specific conditions (e.g., diabetes, cancer).

Health Scores. We leveraged our categories of medical conditions to define health scores with which to estimate the prevalence of different diseases across states in the U.S. Given the set of

[1] https://social-dynamics.net/MedDL

Table 1: Link between health scores computed at the level of U.S. states and official health statistics. Pearson correlations r are reported for $\rho = 0$ (medical conditions with equal weighting), $\rho = 1$ (medical conditions weighted by their centrality on the co-occurrence graph). P-values classes are also reported (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

Health score	Official Statistic	$r_{\rho=0}$	$r_{\rho=1}$
$H^l_{S_i}$	**Mental Health**		
mental	Mentally Unhealthy Days	-.31*	-.45**
mental	Mental Illness	-.23*	-.30*
$H^l_{S_i}$	**Substance Abuse**		
breathing	Cigarette Use	-.31*	-.29*
infections	Cocaine Use	-.25	-.29*
infections	Heroin Use	-.30*	-.43**
$H^l_{S_i}$	**Metabolic syndrome**		
obesity	High Cholesterol Prev.	-.29*	-.46***
obesity	High Blood Pressure	-.26	-.45***
obesity	Mortality Cardiovascular	-.19	-.39**
obesity	Mortality CHD	-.16	-.47***
obesity	Mortality Heart Disease	-.21	-.39**
obesity	Overweight	-.01	-.33*
obesity	Diabetes Prev.	-.25	-.45***
$H^l_{S_i}$	**Specific Diseases**		
elderly	Arthritis	-.45**	-.47***
breathing	Asthma	-.33*	-.42**
$H^l_{S_i}$	**STDs**		
STDs	HIV prevalence	-.23	-.43**
STDs	AIDS prevalence	-.22	-.41**
STDs	Prim. and Sec. Syphilis	-.22	-.47***
STDs	Early Latent Syphilis	-.28*	-.39**
H^l_S	**All Conditions**		
all	Poor Self-rated Health	-.34**	-.33*
$H^l_{S_c}$	**Most Central Conditions**		
most central	Poor Self-rated Health	-.38**	-.39**

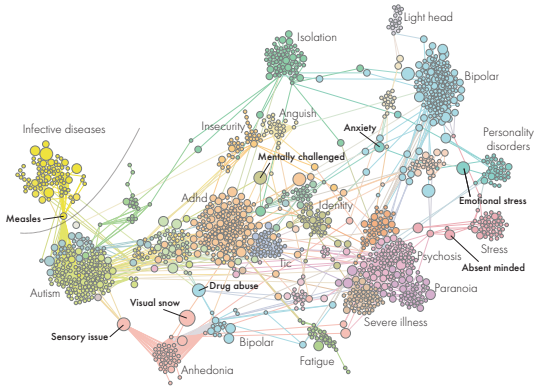

Figure 2: Co-occurrence network of symptoms in the mental health category. Colors represent level-2 categories and node size is proportional to the number of level-1 categories the nodes belong to. The names of some categories are reported. The names of some nodes that belong to multiple categories is reported in bold. The infective diseases category belongs to a different level-1 category.

conditions $S_i \in S$ in category i, and a user u resident in location (state) l, we considered the set of conditions $S_i(u) \in S_i$ that user u has mentioned. We then computed the weighted fraction of users in location l who mentioned any condition of category i :

$$f_i^\rho(l) = \frac{1}{|U_l|}\left(\sum_{u \in U_l} (\max(\{c_{pr}(s), \forall s \in S_i(u)\})^\rho\right) \tag{1}$$

where U_l is the set of all users in state l, $S_i(u)$ is the set of conditions in category i that user u mentioned, $c_{pr}(s)$ is the Page Rank centrality of condition s (Page et al., 1999), and ρ is equal to either zero or one depending on whether Page Rank centrality is used or not. When $\rho = 0$, the centrality value is discarded, and $f_i^{\rho=0}(l)$ becomes simply the fraction of users in l who mentioned conditions in category i. By weighting conditions by their centrality on the co-occurrence network, we wanted to give more importance to those conditions that best represent the category they are in. Given these (weighted) fractions, we computed a health score H_i^l for category i at location l:

$$H_i^l = -\frac{(f_i^\rho(l) - \mu_i)}{\sigma_i} \tag{2}$$

where μ_i and σ_i are the mean and standard deviation of f_i^ρ across all locations. The minus makes it possible to have positive values of H_i^l representing "healthy" areas where the fraction of people mentioning conditions in i was lower than average, and negative values representing areas where those symptoms were mentioned more frequently than expected (Kramer, 2010; Bagroy et al., 2017).

We computed three types of health scores: H^l_S, $H^l_{S_i}$, and H^l_{Sc}. We defined them based on three sets of medical conditions: the full set of conditions from all categories (S); the conditions in category i (S_i); and the set of most-central conditions in the co-occurrence network (top 5% in the PageRank distribution) (S_c).

3 RESULTS

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 (Appendix Table 2). 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. We collected official data from the Centers for Disease Control and Prevention (CDC)[2] and from the Substance Abuse and Mental Health Services Administration (SAMHSA),[3] both of which regularly publish health statistics in the U.S. 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). All SAMHSA statistics were compiled between 2017 and 2018. In total, we collected 18 health statistics ("Official statistic" column in Table 1).

We tested two hypotheses: **(H1)** the prevalence of a specific health condition i measured by official statistics negatively correlates with the corresponding health score H_i^l; and **(H2)** poor self-reported general health negatively correlates with our general health scores H_S^l and $H_{S_c}^l$.

The correlation results summarized in Table 1 suggest that our indices are able to capture real-world prevalence at varying degrees, and that the knowledge of the structure of the co-occurrence network is useful. The health scores computed with such a knowledge ($\rho = 1$) yield stronger and more significant correlations compared to no knowledge, which correspond to mention counts ($\rho = 0$). Indeed, the centrality-weighted scores achieve strong correlations with statistics on specific conditions ($-0.29 \leq r \leq -0.47$, which supports **H1**), and with the statistic on overall poor health ($-0.33 \leq r \leq -0.39$, which supports **H2**). 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 on understanding of 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. The taxonomy enables qualitative investigations, too. For instance, doctors could study connections between subclasses of mental diseases (Figure 2) for unexpected links. One such an example is the link between infective and mental disease clusters through 'measles' revealing people's perception.

4 DISCUSSION

Upon extracting medical mentions from a large social media corpus, and examining the network of their co-occurrences, we derived the first comprehensive online health taxonomy. The emergent categories in this taxonomy align well with the official disease categorization. Furthermore, our health scores computed from Reddit strongly correlated with the prevalence of their corresponding diseases at the level of U.S. states. One of the main limitations of our approach is the restricted number of data points for our correlation analyses, which were bounded by the number of states in the U.S. Repeating our study on geo-referenced data with a finer spatial granularity would partly address this issue. Second, this study aims at providing broad and robust health indices that could be integrated with results from official surveys, potentially providing semantics on top of pure prevalence estimates. More work should go into how such an integration could be achieved. For example, one could explore the applicability (and the potential limitations) of existing statistical methodologies that aim at combining survey data with social media indicators (Alexander et al., 2020).

[2] `https://www.cdc.gov/DataStatistics`

[3] `https://www.samhsa.gov/data`

REFERENCES

Monica Alexander, Kivan Polimis, and Emilio Zagheni. Combining social media and survey data to nowcast migrant stocks in the united states. *arXiv preprint arXiv:2003.02895*, 2020.

Shrey Bagroy, Ponnurangam Kumaraguru, and Munmun De Choudhury. A social media based index of mental well-being in college campuses. In *Proceedings of the Conference on Human factors in Computing Systems (CHI)*, pp. 1634–1646, 2017.

Duilio Balsamo, Paolo Bajardi, and André Panisson. Firsthand opiates abuse on social media: Monitoring geospatial patterns of interest through a digital cohort. In *Proceedings of the ACM World Wide Web Conference (WWW)*, pp. 2572–2579, 2019.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In Munmun De Choudhury, Rumi Chunara, Aron Culotta, and Brooke Foucault Welles (eds.), *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pp. 830–839. AAAI Press, 2020.

Michele Coscia and Frank MH Neffke. Network backboning with noisy data. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pp. 425–436. IEEE, 2017.

George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. Characterisation of mental health conditions in social media using informed deep learning. *Scientific Reports*, 7:45141, 2017.

Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991, 2015.

Adam DI Kramer. An unobtrusive behavioral model of gross national happiness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 287–290. ACM, 2010.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

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.

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)*, pp. 170–181, 2020.

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

A APPENDIX

(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, only, spots, bleeding, burns, inflammation, rash, itching, eczema, allergies, bites, herpes, food allergies, soreness, bumps, psoriasis, vitamin, body hair, 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, pain, 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, spasms, severe disfunctions	hard skin, spasm, fainting, arrhythmia
digestive [13]	bloating, chron, hemorroid, irritation, bowel inflammation, celiac, constipation, stomach, diarrhea, gastritis, flu	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 2: (*A*) The taxonomy of medical conditions extracted from Reddit, arranged in two levels, with some examples of individual entities. The names of the level-1 and level-2 categories were assigned by the authors after manual inspection. We manually arranged the top-level categories into six coherent themes. The numbers next to the level-1 category names correspond to the matched ICD-11 categories. (*B*) The list of top-level categories from International Classification of Diseases (ICD-11) by the World Health Organisation (WHO).