Talking about diseases: A survey of patient-prioritised disease phenotypes

Luke Slater
University of Birmingham
Slides: https://lokero.xyz/tad.pdf

HDRUK
Health Data Research UK

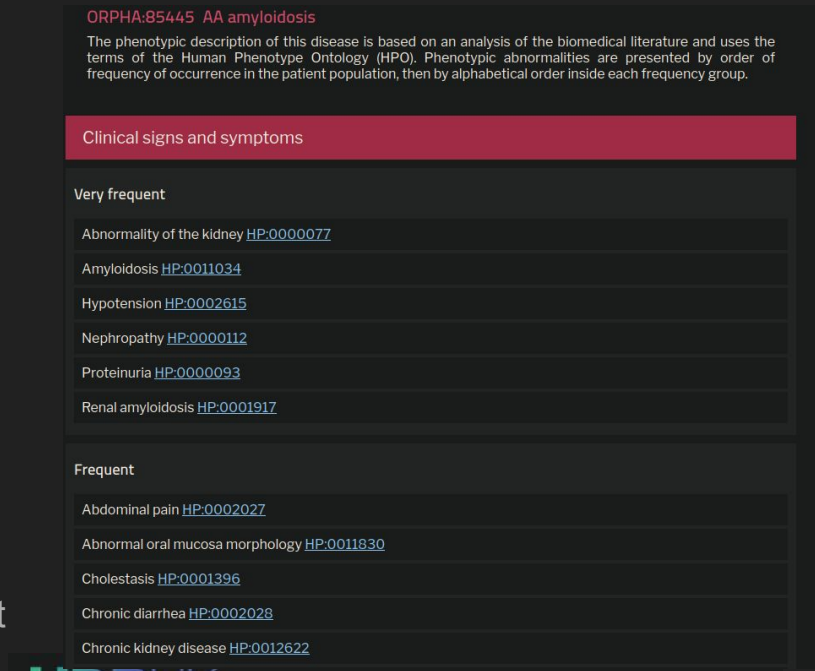
UNIVERSITY OF BIRMINGHAM

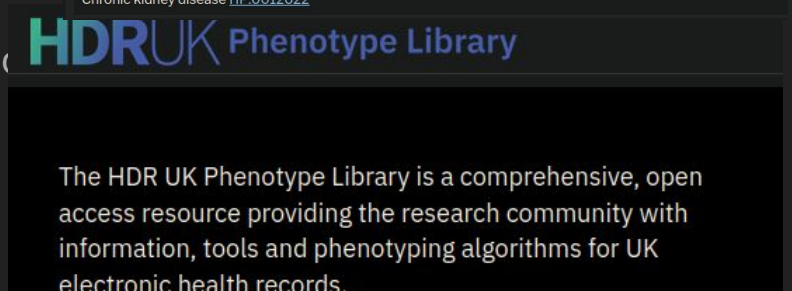
Overview

- Digital Phenotyping
- Aims and Hypotheses
- Exciting methodological challenges
- Methodology: Solutions to exciting methodological challenges
- Social Media Digital Phenotype
- Biomedical and Literature Digital Phenotype
- Analysis etc

Digital Phenotypes and Digital Phenotyping

- Digital phenotypes are representations of the phenotype of some healthcare entity
  - Practice starting mostly with descriptions of rare & genetic diseases (right)
  - They are useful as background knowledge: for personalised medicine, differential diagnosis, clinical awareness, secondary research, hypothesis generation, etc
- Digital phenotyping describes the process of using novel digital information to produce a digital phenotype
  - E.g. tracking social interaction via social media, heart rate on smart watch, etc
  - Recently popular, but mostly applied to describing and tracking individual patients for stratification etc, not feeding back to digital phenotype resources





Understanding Diseases

- Digital phenotypes that do exist for diseases encode institutional understandings of the diseases, usually derived from literature
- This is useful, but we know there is a gulf in understanding about disease, priorities
- Pendleton 2021 and many others showed that patients use different language to talk about diseases
- By understanding the way that patients conceptualise diseases, language that patients talk about their diseases, and the priorities that they have when talking about them, we can more accurately understanding their experience
- Compare patient voice to institutional resources to bridge this gulf through mutual or reciprocal awareness
- This information can inform clinical awareness, as well as support improved analysis e.g. clinical decision support, patient stratification, etc

Herniated disc:

- Sciatica
- Abnormal gait
- Difficulty sleeping

Patient X:

- Lower back pain
- Difficulty sleeping
- Insomnia
- Difficulty walking
- Anxiety

Project Aims and Hypotheses

- Can we identify sensical digital phenotypes, disease-phenotype associations from social media data?
- Can we explore how these differ from institutional digital phenotypes, understandings and priorities?
- Can patient-prioritised phenotype associations reveal underfocused considerations for diseases that can inform clinical awareness or generate hypotheses for research?

Raw Social Media Data

- Data is in the form of ‘transactions’ text mined from social media by White Swan
- Each transaction represents a social media post
- Each transaction is linked with zero or more diseases and phenotypes, representing those that were mentioned in the social media post
- Labels referring to the same phenotype/disease have been linked via an ontology term ID
- 5,842 unique diseases and 6,817 unique phenotypes are mentioned across 138,977,481 transactions

Exciting Methodological Challenges

Normalised Pointwise Mutual Information (NPMI) ia standard way of measuring association between entities in a dataset

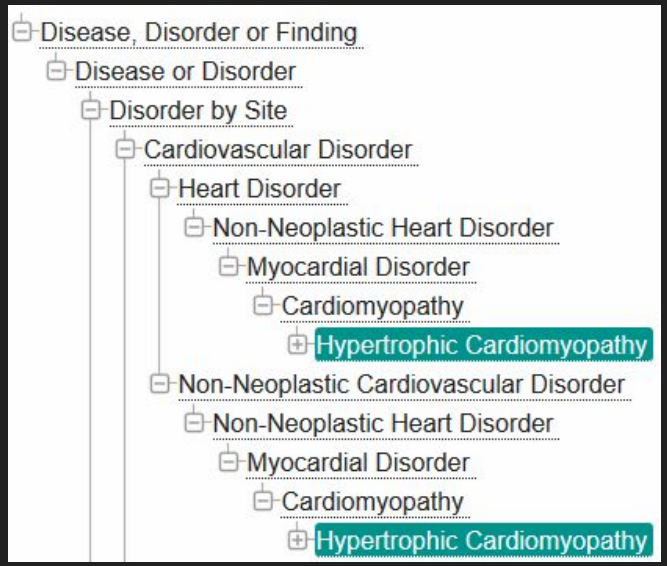
$$PMI(a,b) = \log(\frac{P(a,b)}{P(\mathrm{a})\mathrm{P}(\mathrm{b})})$$

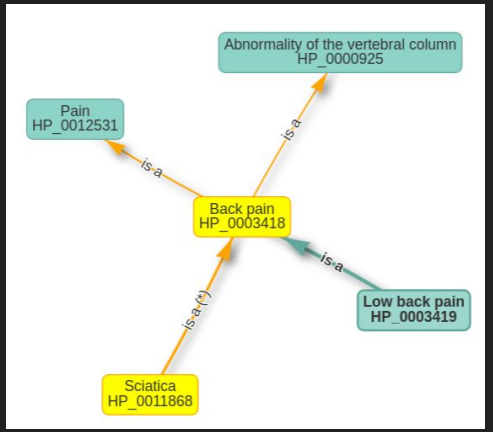
This is great, but there are some problems:

- Ontology-linked data are hierarchical, how do we take this into account?
- At what NPMI level do we determine that a phenotype is associated?
- Results are massively skewed and unhelpful for classes that appear rarely: e.g. if ‘Diastolic heart failure’ appears ~20 times, and ‘bladder polyp’ appears ~10 times, but half of those times are in the same message as diastolic heart failure’, this produces an NPMI of 0.5.
- How do we identify associations that are unique to social media?
- How can we minimise error?
- How can we do all of this on such a massive dataset?

Integrating background knowledge

- Our transactions mention phenotypes and diseases by linking them to ontology classes
- For our purposes, ontologies are simply graphs that describe more and less specific diseases and phenotypes
- When we're talking about sciatica, we're also talking about back pain, and we're also talking about pain
- We used the ontology graph to propagate mentions of diseases and phenotypes across the whole record





Identifying Good Associations

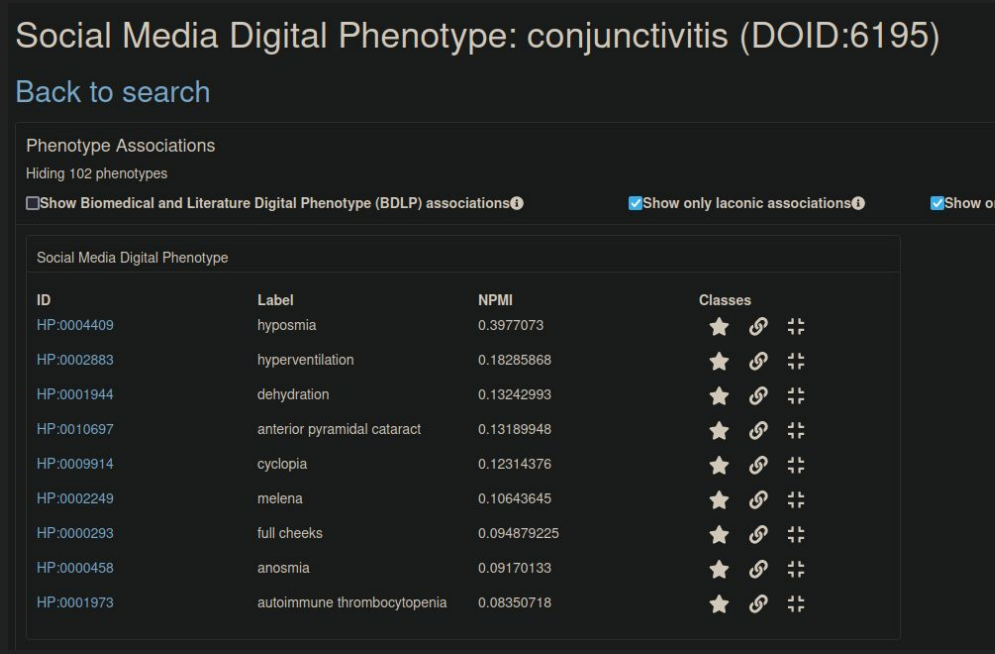
- How can we avoid skewed values from rare diseases and phenotypes, identify cut-offs for ‘significantly’ associated variables, avoid erroneous relationships?
- Other works based on literature use optimisation at a particular task, e.g. variant prediction, but this is not really suitable
- Permutation testing!
- We used results from 1,000 monte carlo simulations on 10% samples to:
  a. Remove phenotypes and diseases with insignificant ocurrence ($p>0.001$)
  b. Calculate q-values for remaining associations with acceptable false discovery rate of 0.005
  c. Save only Significant associations
- The bar is set very high here

Comparing with Institutional Biomedical and Literature Knowledge

- If we want to identify novelty in our social media phenotype associations, we need to collect associations from the institutional sources
- To do this, we integrated three sources
  - Text mining over Pubmed: Collier 2021
  - Text mining over Pubmed: Kafkas 2021
  - Semi-automatic from structured biomedical resources: Kafkas 2021
- These cover a wide range of biomedical and literature resources describing diseases
- Associations have been derived using a related methodology (i.e. NPMI), but with some curious differences
- We call this integrated resource the Biomedical and Literature Digital Phenotype (BDLP)

Dashboard

- There is a web dashboard for exploring the social media digital phenotype: https://smdp.lokero.xyz/
  - Username: test
  - Password: `phenotest`
- Only currently for demonstration; do not consider results final or usable



Results

- We created the Social Media Digital Phenotype (SMDP): a lot of associations
- We matched 355 diseases between the two sets
- Using semantic similarity, we can use the social media derived phenotype associations to predict the disease based on its BLDP (AUC=0.832)
- Across the 355 linked diseases, we identified 11,976 novel significant associations in the social media digital phenotype

Subset	Source	Diseases	Associations	
			All	Significant
Unlinked	Biomedical Database and Literature (BDL-DP)	6,155	2,368,666	-
	Social Medial (SM-DP)	5,562	5,619,427	28,789
Linked	Biomedical Database and Literature (BDL-DP)	355	267,757	-
	Social Media (SM-DP)	355	63,325	25,960

Analysis

- Nervous and digestive system phenotypes are over-represented in novel associations
- 340 of the 355 matched diseases yielded novel nervous system phenotypes
- 112 unique nervous system phenotypes were identified
- We can use Klarigi[1] to identify the composition of the new associations across the 293 diseases (below)
- **Social media analysis reveals novel behavioural & mental health phenotypes for diseases; potentially underfocused in literature despite being a focus for public**

Class	Inclusion	IC
Behavioral abnormality (HP:0000708)	0.88	0.47
Abnormal emotion/affect behavior (HP:0100851)	0.54	0.58
Abnormality of higher mental function (HP:0011446)	0.52	0.5
Reduced consciousness/confusion (HP:0004372)	0.46	0.62
Sleep disturbance (HP:0002360)	0.44	0.59
Abnormality of movement (HP:0100022)	0.44	0.47
Autistic behavior (HP:0000729)	0.41	0.61
Impairment in personality functioning (HP:0031466)	0.38	0.59
Abnormal social behavior (HP:0012433)	0.35	0.6
Impaired social interactions (HP:0000735)	0.34	0.64
Abnormal central motor function (HP:0011442)	0.34	0.47

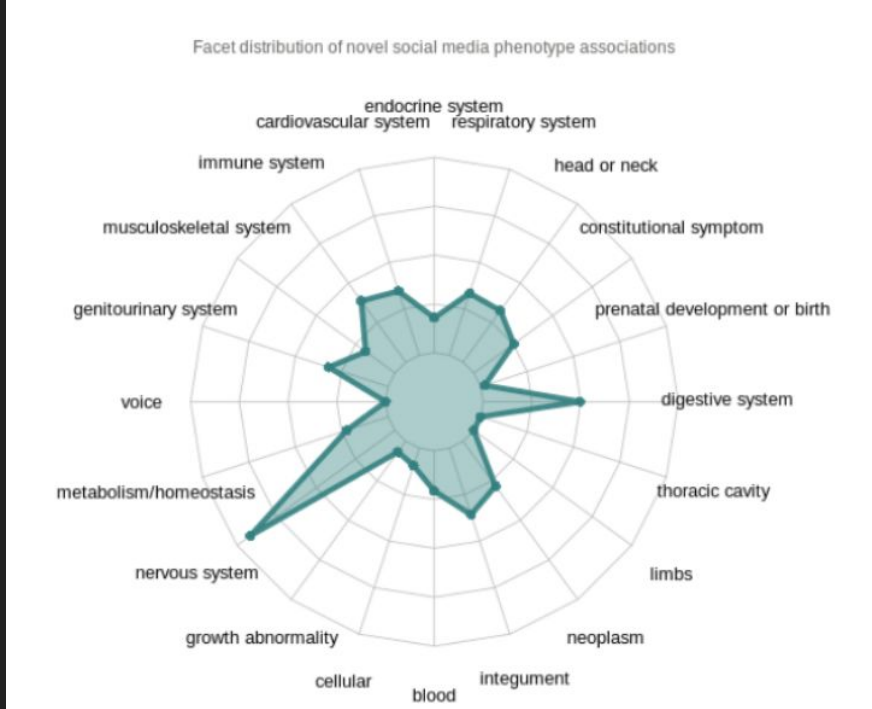
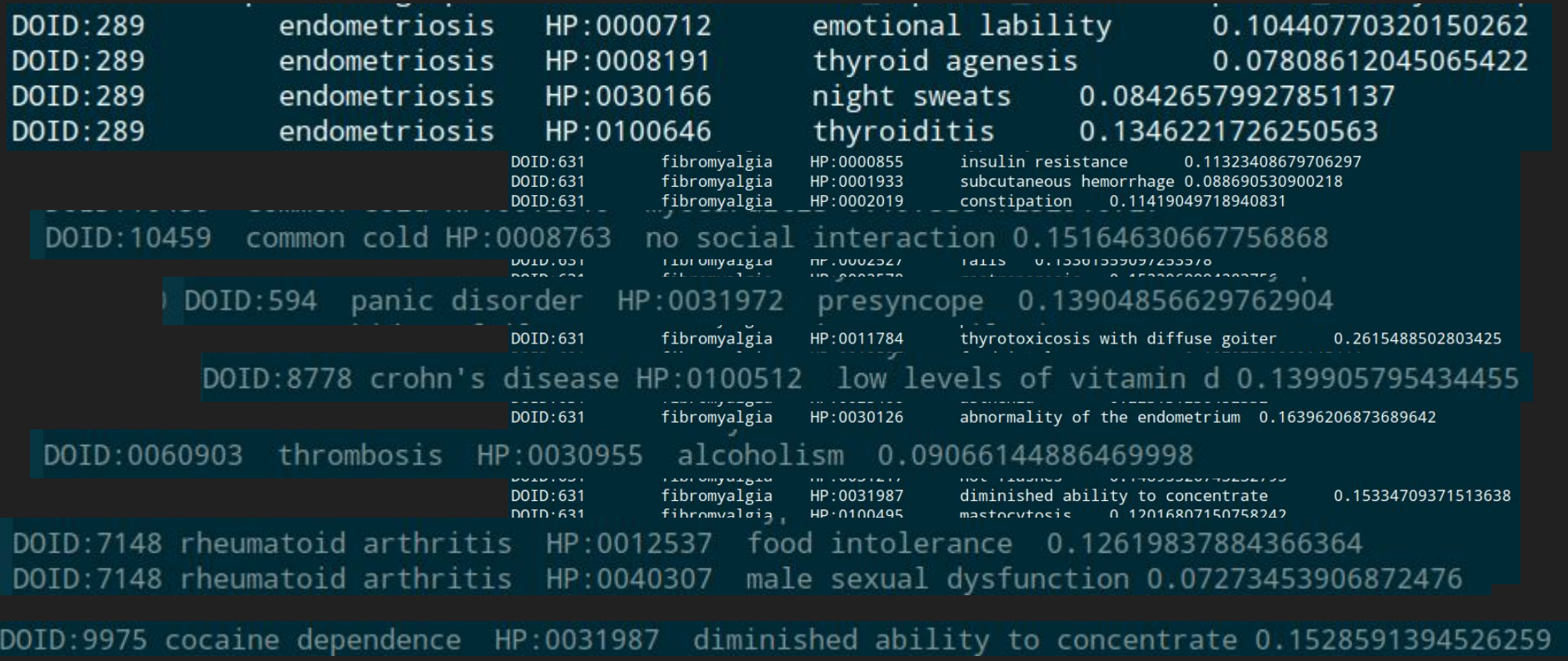


Figure 1. Distribution of phenotype associations across HPO facets for the social media and literature dataset. Values derived from proportion of total phenotypes in each set. max is 0.175

1. https://www.sciencedirect.com/science/article/pii/S0010482522011337

There are many interesting novel associations recovered, including those for diseases with complicated and poorly understood phenotypic profiles:



Social Media Digital Phenotype: covid\-19 (DOID:0080600)

Back to search

Phenotype Associations

Hiding 179 phenotypes

☐ Show Biomedical and Literature Digital Phenotype (BDLP) associations ☑ Show only laconic associations ☐ Sho

Social Media Digital Phenotype

ID	Label	NPMI	Classes
HP:0004469	chronic bronchitis	0.39478168	
HP:0004409	hyposmia	0.34984446	
HP:0025143	chills	0.3432219	
HP:0031246	nonproductive cough	0.3267252	
HP:0002326	transient ischemic attack	0.31728566	
HP:0002098	respiratory distress	0.31267363	
HP:0011949	acute infectious pneumonia	0.30501083	
HP:0002883	hyperventilation	0.2980205	
HP:0000019	urinary hesitancy	0.28245595	
HP:0000458	anosmia	0.27626094	
HP:0002725	systemic lupus erythematosus	0.2746181	
HP:0025095	sneeze	0.27180508	
HP:0002878	respiratory failure	0.2586334	
HP:0031417	rhinorrhea	0.2547417	
HP:0006528	chronic lung disease	0.24593034	
HP:0025439	pharyngitis	0.23489639	

Conclusions, Limitations, and Current Status

- Conclusions:
  - Social Media Digital Phenotype recapitulates and extends biomedical + literature knowledge
  - We can find novel phenotypes with little basis or underfocus in literature
  - Mental health and digestive phenotypes are over-represented in this set
- Limitations:
  - Manner of relationship unknown (problem shared with existing resources)
  - Errors of variability due to different use of words, e.g. plethora
  - we're setting the bar extremely high for the associations we include. For example, the occurrence appearance is unnecessarily ruthless, a more appropriate test would be if these labels were to show up across other random language; also currently only looking binary only
- I am currently writing a manuscript with a focus on this being a survey/resource for clinical awareness of patient priorities
  - I am seeking clinicians to write some interpretation and clinical analysis for particular diseases, maybe one or two interesting case studies
- Future work:
  - Deep dives on particular diseases or disease areas
  - Comparison with clinical phenotype (i.e. clinical letters)
  - Secondary analysis performance (.e.g differential diagnosis)
  - Extract additional value
  - Resource for clinical awareness
- Thanks also to my collaborators in the Gkoutos Lab, Paul Schofield, and White Swan
- Questions