Automatic Extraction of Fatigue Content from Voice Diary Data

Methodological Issue Being Addressed

Is it feasible to extract measures of patient fatigue from multi-lingual voice diary data using a fully automated analytic pipeline, and will these measures correlate with patient reported outcomes?

Introduction

Patient reported outcomes (PRO) are a cornerstone of clinical trials but can be too burdensome for high frequency use. Fatigue is present in many neurological and immune mediated inflammatory disorders and chronic diseases such as IBD and is often the symptom with the greatest patient impact.
A daily voice diary could provide a low-burden method for collecting high-frequency patient fatigue symptom data, but manual transcription and analysis of this data would limit scalability.
We developed an automated analytic pipeline to extract fatigue symptom content from multi-lingual voice diaries. We evaluated whether fatigue symptoms measures derived using this method correlate with validated PRO measures of fatigue.

Methods

The IDEA-FAST consortium is investigating digital biomarkers of fatigue and sleep in neurological and immune disorders. A feasibility study was conducted prior to a larger clinical observational study. This included an optional brief daily voice diary task delivered via an app. Patients with Neurodegenerative Disorders (Parkinson's disease, Huntington's disease), Immune mediated inflammatory disorders (Lupus, Rheumatoid Arthritis, Primary Sjogren's Syndrome, IBD) and healthy controls were recruited (Table 1). Patients were English, Dutch and German speakers and completed daily open-ended voice diary entries describing their day, in their own language, over the course of a four-week study.
Patients also completed the FACIT-F Fatigue Scale weekly.

The voice diary data was automatically transcribed and translated to English using the Whisper ASR model. The translated transcripts were divided into sentences using the Spacy NLP toolkit and each sentence manually labelled by content type, including whether the sentence describes a state of fatigue or subjective energy.

An automatic analytic pipeline was developed employing pre-trained large language models to extract semantic vectors for each sentence.

The sentences "I feel tired" and "I feel full of energy" were chosen as canonical descriptions of fatigue or energy in this dataset, and semantic vectors were calculated for these two reference sentences. For each sentence we calculated the cosine distance to the closest reference sentence and constructed a fatigue score as $(1 - \text{normalised closest reference distance})^2$, multiplied by -1 if the sense was negative (closest to "I feel tired").

For each participant the mean of their weekly FACIT-F score was calculated and correlated with the mean of the semantic fatigue measure.

Candidate pretrained language models were selected based both on literature and ratings on the Huggingface model website. In total 5 models were selected and compared. All-mpnet-base-v2 and all-MiniLM-L6-v2 were selected based on popularity on Huggingface for Sentence Similarity scoring. The all-Roberta-large-v1 model was the best performing model at the time of abstract submission. The xlm-Roberta-large model was selected for its multilingual capability and is the basis for the current best performer: the multilingual-e5-large which has been created by supervised fine-tuning xml-Roberta-large on a labelled dataset.

Table 1: Number of participants per cohort and category

Cohort	Count	Category
Healthy controls	9	Healthy
Inflammatory bowel disease	7	Immunology
Primary Sjogren's Syndrome	7	Immunology
Rheumatoid Arthritis	13	Immunology
Systemic lupus erythematosus	12	Immunology
Huntington's disease	5	Neurology
Parkinson's disease	7	Neurology

Results

64 participants opted to record voice diary entries, producing 853 diary entries, consisting of 5.96 hours of voice audio data. The transcription, translation and tokenisation pipeline yielded 3980 sentences, of which 386 were manually coded as being either fatigue or energy related.

The processing pipeline was run over each of the candidate language models. The best performing language model was the recently release multilingual-e5-large. Using this model the automated pipeline corresponded with PRO measures with a correlation coefficient of .44 as shown in Fig 1.

Fig 1: Correlation between aggregated fatigue scores and FACIT-F

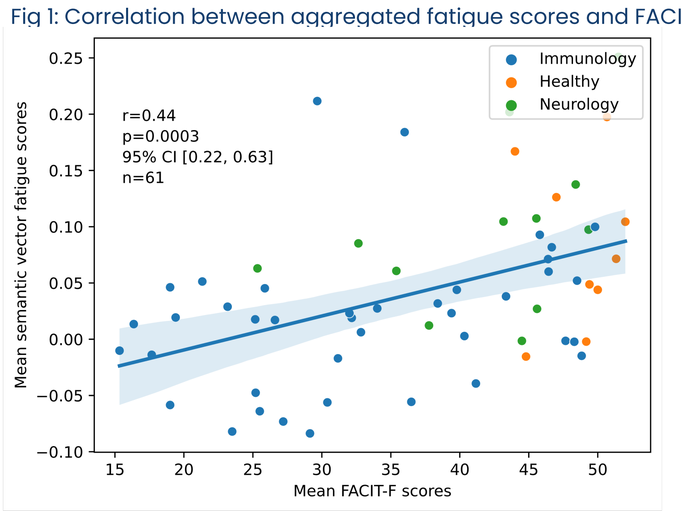
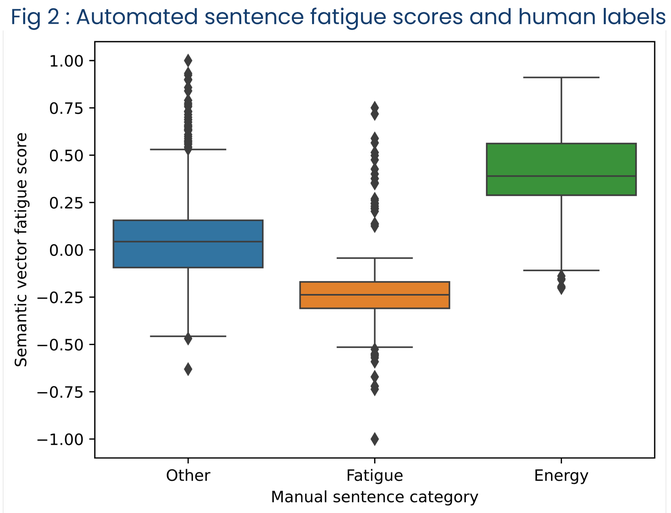


Fig 2 : Automated sentence fatigue scores and human labels



Performance of pretrained language models

Table 2: Properties and performance of the pretrained language models

Model Name	Year	Size	Correlation Coefficient	p-value	Lower 95th CI	Upper 95th CI
all-mpnet-base-v2[1]	2021	438MB	.2503	.0517	-0.0016	0.4723
all-MiniLM-L6-v2[1]	2021	91MB	.2188	.0902	-0.0349	0.4460
all-roberta-large-v1[1]	2021	1.42GB	.2731	.0332	0.0228	0.4912
xlm-roberta-large[2]	2019	2.24GB	.0311	.8119	-0.2224	0.2807
multilingual-e5-large[3]	2023	2.24GB	.4438	.0003	0.2161	0.6257

It is interesting to note that xlm-Roberta-large and multilingual-e5-large models have identical size and architecture and differ only by the additional training of the latter.

Conclusion

These data demonstrate the feasibility of automatically extracting patient symptoms from unstructured speech. These symptom scores correlate with validated scale measures. The multilingual pipeline can extract symptom scores without reliance on human transcription or translation, potentially enabling deployment at scale. Further analytical work will examine the validity of this finding by exploring other reference sentences and address reliability by analysing data at multiple timepoints.

Authors & contacts

Nick Taptiklis[1], Francesca Cormack[1], Michele Veldsman[1], Zsolt Homoridi[2], Teemu Ahmaniemi[2], Walter Maetzler[3], Fai Ng[4]

[1]Cambridge Cognition, United Kingdom
[2]VTT Technical Research Center of Finland Ltd., Espoo, Finland,
[3]Department of Neurology, University Medical Center Schleswig-Holstein Campus Kiel, Germany
[4]Translational and Clinical Research Institute, Newcastle University and NIHR Newcastle Biomedical Research Centre, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK

References

1. Developed during the Huggingface "Community week using JAX/Flax for NLP & CV"
2. A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale," arXiv preprint arXiv:1911.02116, 2020.
3. L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei, "Text Embeddings by Weakly-Supervised Contrastive Pre-training," arXiv preprint arXiv:2212.03533, 2022.

IDEA FAST

CAMBRIDGE COGNITION