Catching the wave: Neuro-symbolic approaches help manage medical evidence

Prof. Dr. Janna Hastings

Medical Knowledge and Decision Support

Institute for Implementation Science in Health Care, Faculty of Medicine, University of Zurich

School of Medicine, University of St. Gallen

janna.hastings@uzh.ch

@jannahastings.bsky.social

https://hastingslab.org/

Illustration: Sara Gironi Carnevale (Science Magazine)

28 February 2025 – St. George's, London

Illustration: Sara Gironi Carnevale (Science Magazine)

What makes it so hard to manage the medical evidence base?

COMMENT | 15 December 2021

Decision makers need constantly updated evidence synthesis [1]

Fund and use 'living' reviews of the latest data to steer research, practice and policy.

Julian Elliott, Rebecca Lawrence, Jan C. Minx, Olufemi T. Oladapo, Philippe Ravaud, Britta Tendal Jeppesen, James Thomas, Tari Turner, Per Olav Vandvik & Jeremy M. Grimshaw

PubMed®

1,617,233 articles added to PubMed in 2024 [2]

~ 4,431 per day

The mean estimated time to complete [a] review was 67.3 weeks (IQR=42). [3]

~ 1 1/4 years

[1] Nature 600, 383-385 (2021)
[2] https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html
[3] Borah R, Brown AW, Capers PL, et al. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017;7:e012545

How many papers are published each month?

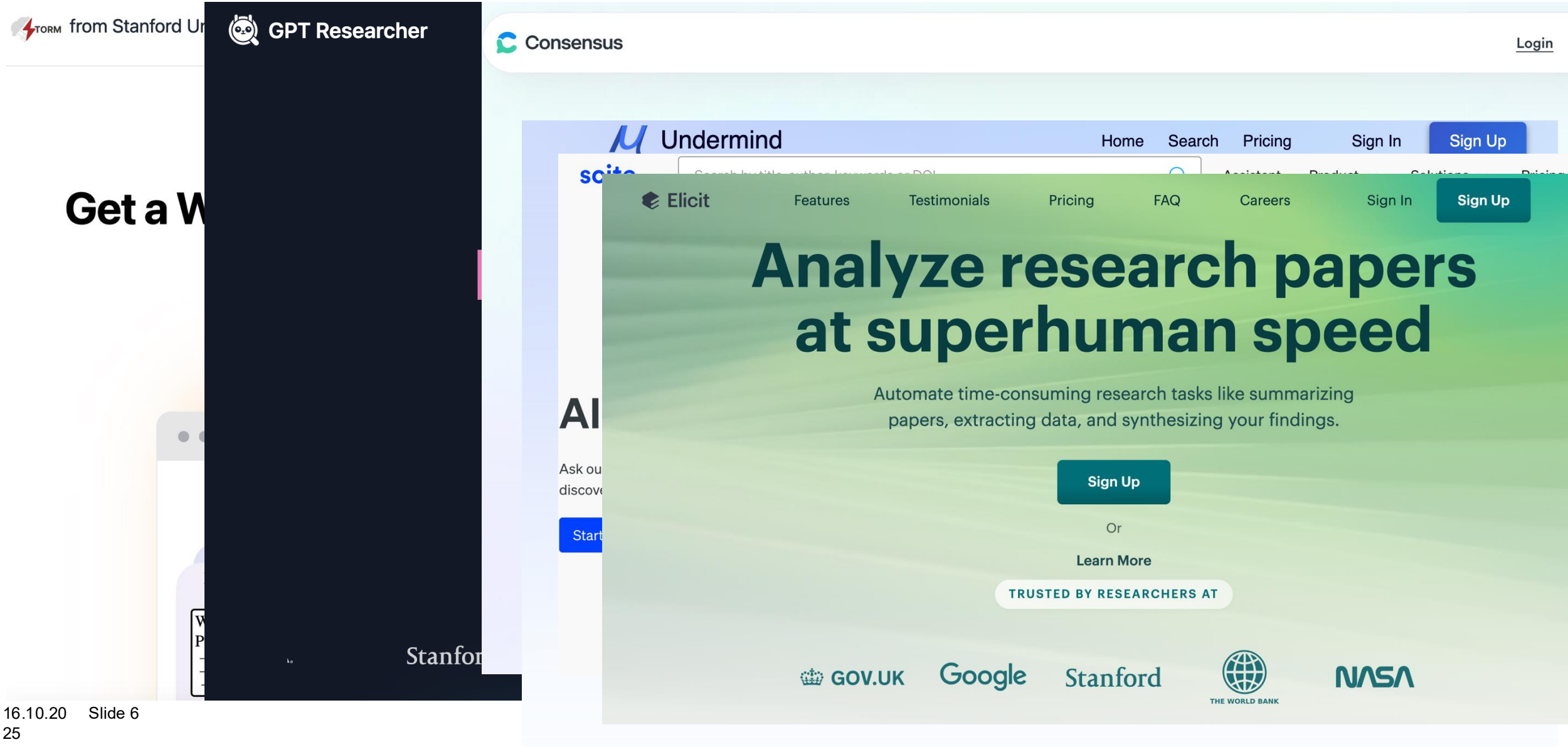
AUTOMATED APPROACHES ARE ESSENTIAL

Large Language Models : the emerging frontier of evidence automation

Image generated with Stable Diffusion (SD3 medium)

- Prior to 2022, no automated system existed with sufficient mastery of human languages to support medical knowledge management, patient query responses etc.
- The breakthrough advance in capabilities in AI-based large language models (LLMs) was due to congruent advances in model architectures, training methods, hardware and data availability
- Now we have many alternative models and tools with similar capabilities, both commercial and openly available

Many LLM-based tools already claim to accelerate evidence reviews



However, these tools are not a replacement for systematic reviews

Feature

CAN AI REVIEW THE SCIENTIFIC LITERATURE?

Artificial intelligence could help to make sense of the world's science — but it comes with risks. **By Helen Pearson**

Faster ?

Poorer quality ?

Some of the newer AI-powered science search engines can already help people to produce narrative literature reviews — a written tour of studies — by finding, sorting and summarizing publications. But they can't yet produce a high-quality review by themselves. The toughest challenge of all is the 'gold-standard' systematic review, which involves stringent procedures to search and assess papers, and often a meta-analysis to synthesize the results. Most researchers agree that these are a long way from being fully automated. "I'm sure we'll eventually get there," says Paul Glasziou, a specialist in evidence and systematic reviews at Bond University in Gold Coast, Australia. "I just can't tell you whether that's 10 years away or 100 years away."

At the same time, however, researchers fear that AI tools could lead to more sloppy, inaccurate or misleading reviews polluting the literature. "The worry is that all the decades of research on how to do good evidence synthesis starts to be undermined," says James Thomas, who studies evidence synthesis at University College London.

Testing LLMs in practice screening studies based on title and abstract

Research | Open access | Published: 15 June 2024

Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain

Fabio Dennstädt, Johannes Zink, Paul Martin Putora, Janna Hastings & Nikola Cihoric

Systematic Reviews **13**, Article number: 158 (2024) | Cite this article

4176 Accesses | **3** Citations | **3** Altmetric | Metrics

Dr. Fabio Dennstädt
University of St. Gallen
and Inselspital Bern

Abstract

Background

Systematically screening published literature to determine the relevant publications to synthesize in a review is a time-consuming and difficult task. Large language models (LLMs) are an emerging technology with promising capabilities for the automation of language-related tasks that may be useful for such a purpose.

Methods

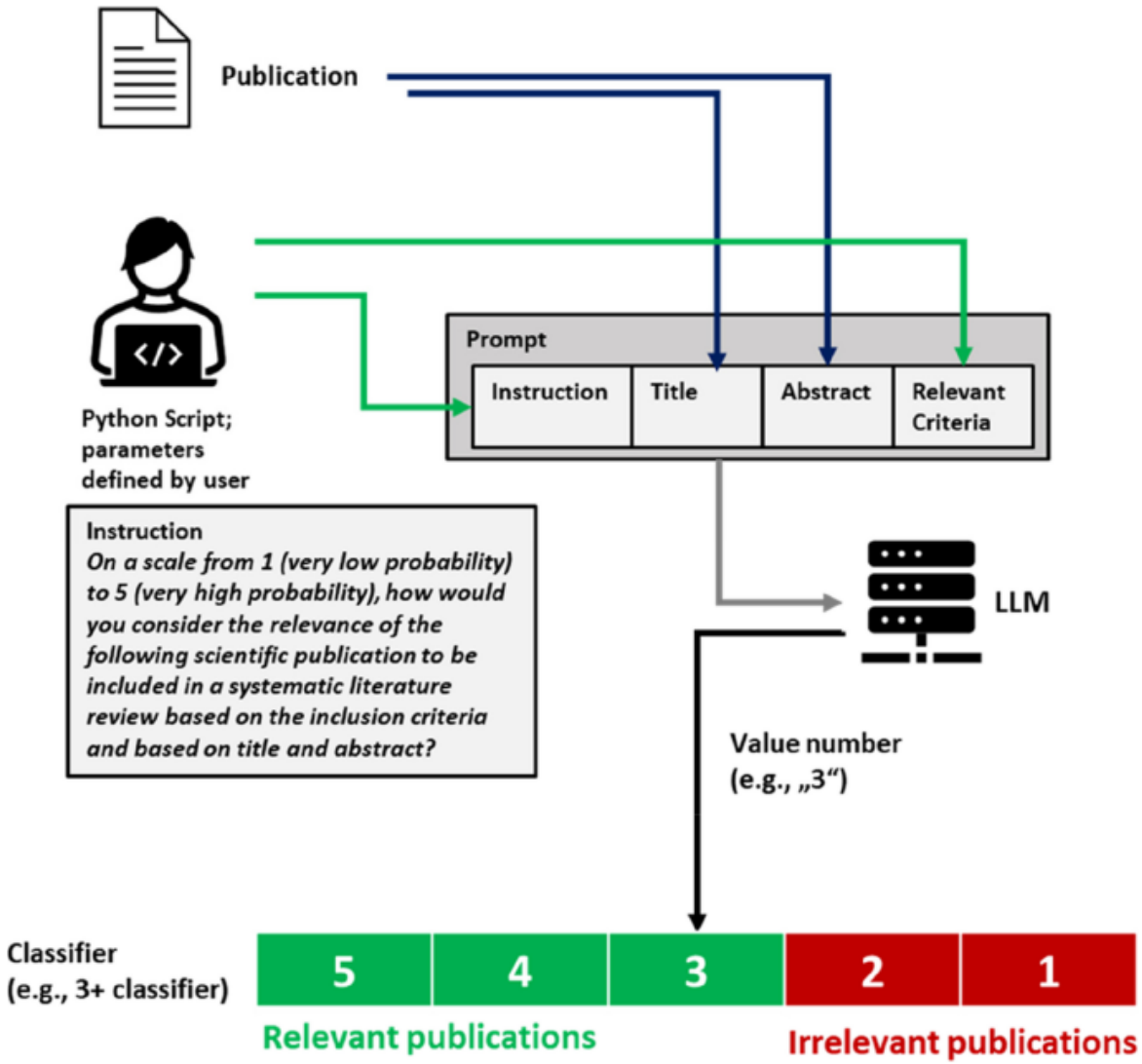
LLMs were used as part of an automated system to evaluate the relevance of publications

Approach: Single-prompt screening task

Title and abstract

Inclusion criteria

Flexible threshold



Model gives rating

Dennstädt et al., Systematic Reviews 2024 **13**, 158

Strategy for evaluation

Open-Source Language Models

for reproducibility and a controlled execution environment – as requested by a reviewer

- Flan T5 – XXL (Google)
- OpenHermes 2.5
- Mixtral 8x7b-0.1
- Platypus 2 70b Instruct (based on Llama 2)

Open Datasets (from published systematic reviews)

Name	Topic	Number of publications	Relevant publications (%)	Reference
Appenzeller-Herzog_2020	Wilson's Disease	3479	26 (0.75%)	[40]
Bos_2018	Cerebral small vessel disease and dementia	5756	10 (0.18%)	[41]
Donners_2021	Emicizumab	660	15 (2.27%)	[42]
Jeyaraman_2021	Osteoarthritis	1194	96 (2.26%)	[43]
Leenaars_2020	Rheumatoid arthritis	9543	792 (8.30%)	[44]
Mejboom_2021	TNFα-inhibitors and biosimilars	2224	37 (1.66%)	[45]
Muthu_2021	Spine surgery	3254	354 (10.88%)	[46]
Oud_2018	Borderline personality disorder	1053	20 (1.90%)	[47]
van_de_Schoot_2018	PTSD	6225	38 (0.61%)	[48]
Wolters_2018	Dementia and heart disease	5038	19 (0.38%)	[49]

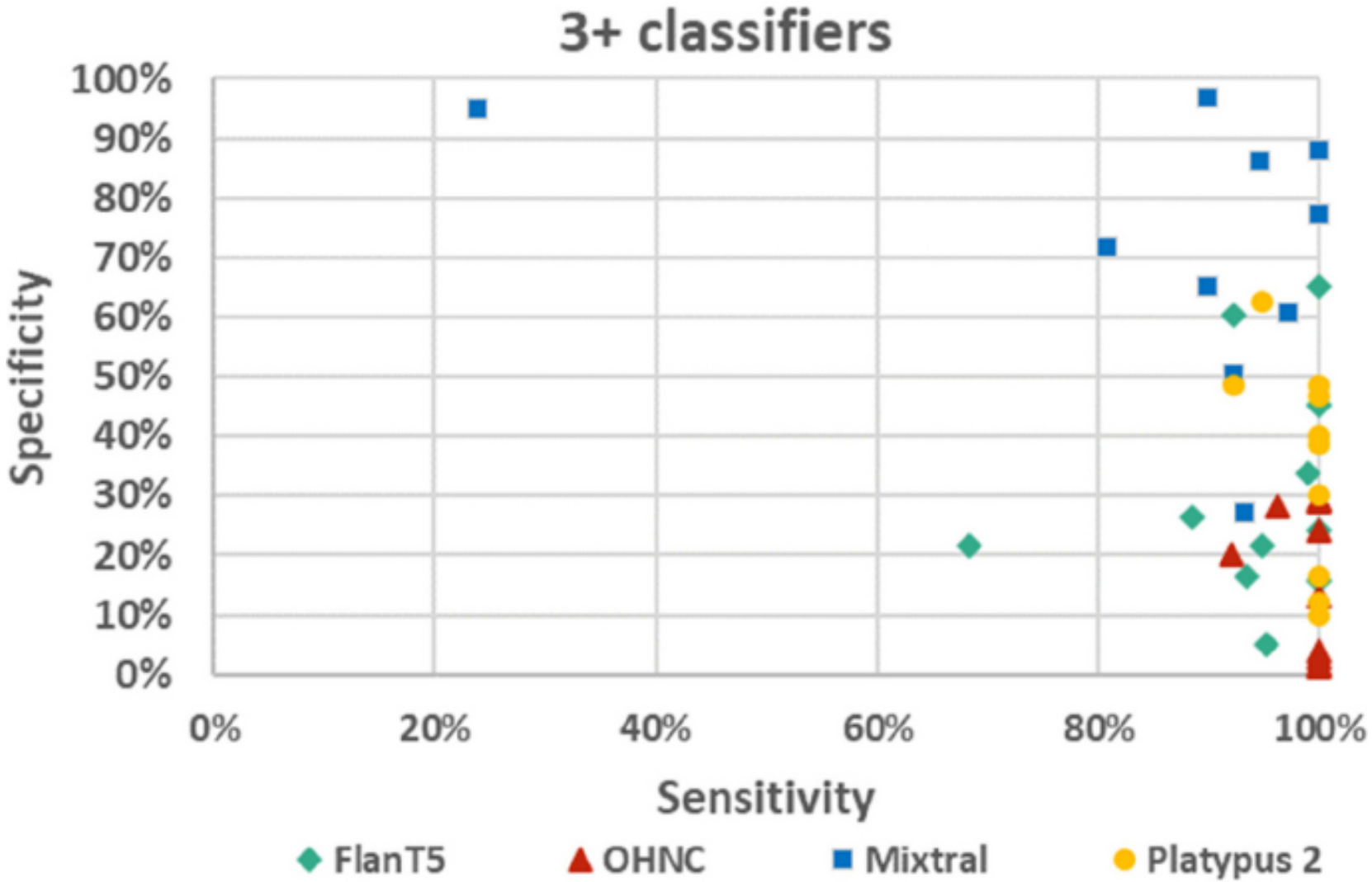
+ 1 newly created dataset

with topic CDS in radiation oncology

521 search results / 36 relevant

Dennstädt et al., Systematic Reviews 2024 **13**, 158

Results: Performance varies by model and dataset



Best-performing model was Mixtral on average, but it still performed poorly (low sensitivity) on one dataset

Nature of the screening criterion

Dennstädt et al., Systematic Reviews 2024 **13**, 158

An aside – Screening is harder than it seems – Is yoga exercise?

Mechanisms through which exercise reduces symptom severity and/or functional impairment in posttraumatic stress disorder (PTSD): Protocol for a living systematic review of human and non-human studies

[version 1; peer review: 1 approved with reservations]

Simonne Wright, Toshi A. Furukawa, Malcolm Macleod, Ouma Simple, Olufisayo Elugbadebo, Virginia Chiocchia, Claire Friedrich, Edoardo G. Ostinelli, Jennifer Potts, Fiona J. Ramage, Spyridon Siafis, Claire Stainsfield, Francesca Tinsdeall, James Thomas, Andrea Cipriani, Georgia Salanti, Soraya Seedat, the GALENOS team

This article is included in The Global Alliance for Living Evidence on aNxiety, depressiOn and pSychosis (GALENOS) gateway

ARTICLE Al

Abstract

Background

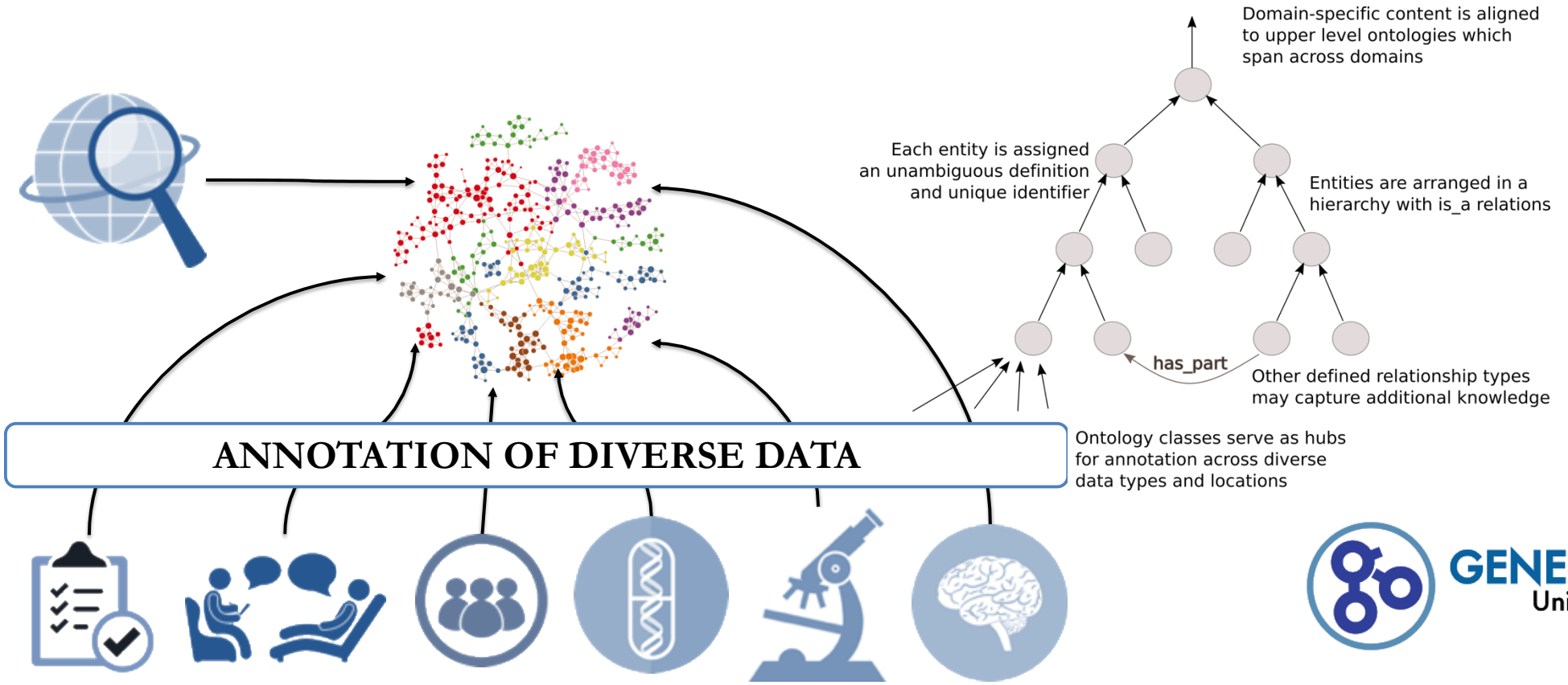
Exercise can play an importan
posttraumatic stress disorder (
management are not fully und
human and human studies to
severity and functional impairn

Table 2. Study inclusion and exclusion criteria for human studies.

Experimental interventions/ exposures	Exercise will include anaerobic exercises (e.g., resistance and strength training, plyometrics), aerobic exercises, exergaming, or active stretching. The exercise may occur before, during, or after the psychotherapy. The exercise can be of any type, length, and strength; it has to be structured and repetitive to differentiate it from other types of physical activities. *We will exclude:* We will exclude Pilates, martial arts, yoga, tai chi, and horseback riding because of the potential psychological benefits of these activities. Inactive stretching will also be excluded.

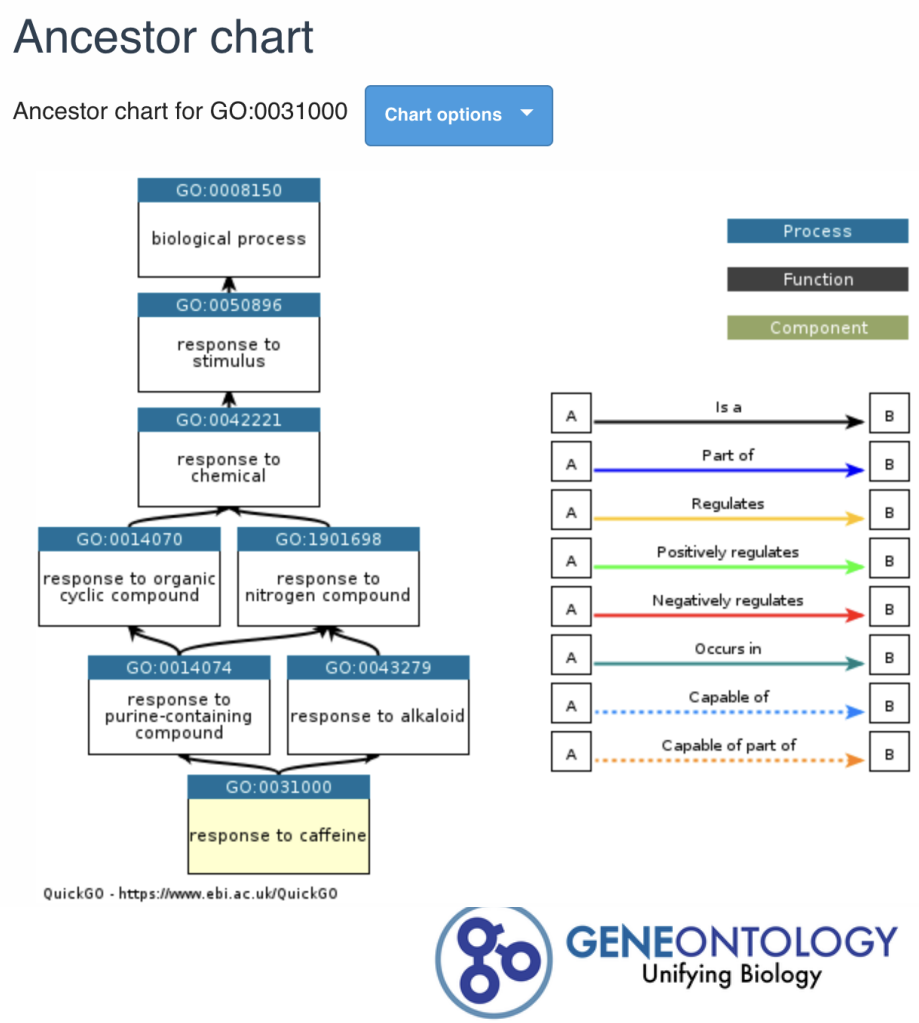
Neuro-symbolic AI approaches – combining ontologies and ML

At least some of the difficulty in gaining an overview of the evidence stems from challenges in agreeing on definitions. **Ontologies** are systems in which agreed-upon definitions can be structured, shared and re-used



Hastings (2017) Methods Mol Biol. 1446:3-13.
Primer on Ontologies.

Ontologies define things (1): – Molecules and their metabolism



GENEONTOLOGY
Unifying Biology

Hastings et al. ChEBI in 2016… *Nucleic Acids Research*, 44 (D1) D1214–D1219,
Neuhaus & Hastings, Applied Ontology 2022 vol. 17, no. 4, pp. 495-513,
Ontology Development is consensus creation, not (merely) representation

Universität Zürich[UZH]
Universität St.Gallen
School of Medicine

The Human Behaviour-Change Project and APRICOT

https://www.humanbehaviourchange.org/

Ontologies define things (2): – What is social support?

social support BCT

https://bciosearch.org/BCIO_007028 Copy Link

ID	Curation status	Created	Modified
BCIO:007028	Published	12 May '23	13 May '24

Parents

entity ≻ occurrent ≻ process ≻ planned process ≻ behaviour change technique ≻ social support BCT

Children

- arrange support BCT
  - arrange emotional support BCT
  - arrange appraisal support BCT
  - arrange informational support BCT
  - arrange instrumental support BCT
- deliver support BCT
  - deliver appraisal support BCT
  - deliver emotional support BCT
  - deliver instrumental support BCT
  - deliver informational support BCT
- advise to seek support BCT
  - advise to seek appraisal support BCT
  - advise to seek emotional support BCT
  - advise to seek informational support BCT
  - advise to seek instrumental support BCT

Definition

A <behaviour change technique> that involves taking steps to secure or deliver the support or aid of another person.

How are ontologies used in evidence synthesis?

Table 1. Baseline characteristics of the trial participants.

	Total ($N = 572$)	CASC ($N = 285$)	Control ($N = 287$)
Age	46 ± 11	47 ± 11	45 ± 12
Age ranges			
<40 years	171 (30)	75 (26)	96 (33)
40–49 years	179 (31)	97 (34)	82 (29)
50–59 years	159 (28)	82 (29)	77 (27)
≥60 years	63 (11)	31 (11)	32 (11)
Male sex	426 (75)	216 (76)	210 (73)
Smoking history			
Years of smoking	25 (18–33)	25 (19–32)	24 (17–33)
Cigarettes per day	20 (15–20)	20 (15–20)	20 (15–20)
Pack-years	20 (14–30)	21 (14–30)	20 (13–31)
Exhaled CO (ppm)	17 ± 11	17 ± 10	18 ± 11
TDS score	7.7 ± 1.5	7.7 ± 1.4	7.8 ± 1.5
FTND	5.3 ± 2.1	5.2 ± 2.0	5.3 ± 2.1
Comorbidities			
Cardiovascular diseases	93 (16)	45 (16	
Respiratory diseases	93 (16)	48 (17	
Psychiatric diseases	31 (5)	12 (4)	
Prescribed Medication			
Varenicline	454 (79)	227 (8	
Nicotine patch	114 (20)	56 (20	
No medication	4 (1)	2 (1)	

Data include mean ± standard deviation, number (% quartile range) scores.
CO carbon monoxide, *FTND* Fagerström test for nicoti
TDS tobacco dependence screener.

In conclusion, a novel digital therapy, the CASC system, significantly improved a long-term CAR from weeks 9 to 24 in conjunction with standard smoking cessation treatment in patients with nicotine dependence. Digital therapy for smoking cessation may be a promising st
prevalence worldwide, and future r
warranted.

METHODS

Study design

This was a multi-center, randomized, controlled, open-label trial (Fig. 2). A detailed study protocol was presented elsewhere[28]. Briefly, participants were recruited from 31 smoking cessation clinics in Japan, from October 2017 to January 2018, and allocated 1:1 to the CASC intervention group and the control group. The intervention group used the CASC system, and the control group used the control app, each for 24 weeks, in addition to a 12-week standard smoking cessation treatment. In reference to the varenicline maintenance therapy for 24 weeks[22], we were interested in whether or not the effectiveness of the CASC system could be maintained after discontinuation of using the app, and we limited access to the app to 24 weeks and evaluated CARs up to 52 weeks. The study was performed in

6

novel smartphone system for smoking cessatio
mobile exhaled CO checker to supplement s
with face-to-face behavioral support and phar
included a long-term observation period (24 ar
measured biochemically validated CARs; and (
additional clinical and psychological efficacy
pharmacotherapy when supplementing a
cessation program. The study also had son
participating institutions were mostly restricted
Japan, where smartphone usage rates were hig
the participants were middle-aged, and the s

Table 2. Changes in MPSS, FTCQ-12, and KTSND scores from baseline to weeks 24 and 52, adjusted by covariates.

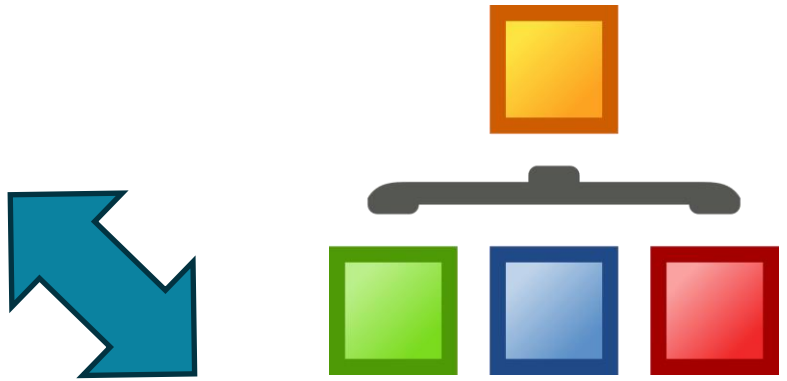
	CASC ($N = 285$)	Control ($N = 287$)	*P*- value[a]
Change from weeks 0 to 24			
MPSS urge total	−1.87 [−2.01 to −1.72]	−1.73 [−1.88 to −1.59]	0.009
MPSS total excluding urges	−0.59 [−0.72 to −0.47]	−0.42 [−0.56 to −0.29]	<0.001
FTCQ-12 emotionality	−1.70 [−1.89 to −1.51]	−1.32 [−1.49 to −1.14]	<0.001
FTCQ-12 expectancy	−2.46 [−2.68 to −2.25]	−2.16 [−2.37 to −1.95]	0.002
FTCQ-12 compulsivity	−1.74 [−1.94 to −1.54]	−1.74 [−1.93 to −1.55]	0.202
FTCQ-12 purposefulness	−2.84 [−3.10 to −2.58]	−2.17 [−2.44 to −1.90]	<0.001
FTCQ-12 general craving score	−2.09 [−2.25 to −1.93]	−1.78 [−1.93 to −1.63]	<0.001
KTSND	−7.0 [−7.7 to −6.2]	−3.9 [−4.5 to −3.2]	<0.001
Change from weeks 0 to 52			
MPSS urge total	−1.82 [−1.98 to −1.67]	−1.65 [−1.81 to −1.49]	0.007
MPSS total excluding urges	−0.52 [−0.64 to −0.39]	−0.42 [−0.54 to −0.29]	0.061
FTCQ-12 emotionality	−1.60 [−1.80 to −1.41]	−1.21 [−1.39 to −1.03]	0.001
FTCQ-12 expectancy	−2.39 [−2.60 to −2.19]	−2.10 [−2.33 to −1.88]	0.002
FTCQ-12 compulsivity	−1.71 [−1.90 to −1.52]	−1.55 [−1.75 to −1.35]	0.019
FTCQ-12 purposefulness	−2.84 [−3.10 to −2.58]	−2.00 [−2.28 to −1.73]	<0.001
FTCQ-12 general craving score	−2.03 [−2.19 to −1.87]	−1.65 [−1.81 to −1.48]	<0.001
KTSND	−5.9 [−6.6 to −5.3]	−4.1 [−4.8 to −3.5]	<0.001

Mean [95% CI] scores are provided. Covariates: medications (varenicline, nicotine patch, or none) and medical institutions.
MPSS Mood and Physical Symptoms Scale (ranging from 0 to 5), *FTCQ-12* French version of the Tobacco Craving Questionnaire 12 (ranging from 1 to 7), *KTSND* Kano Test for Social Nicotine Dependence (ranging from 0 to 30), *CI* confidence interval.
[a]Analysis was based on analysis of covariance.

24 weeks that obtained an additional effect of 12% on the CAR compared

4. Ikeda, N. et al. What has made the population of Japan healthy? *Lancet* **378**,

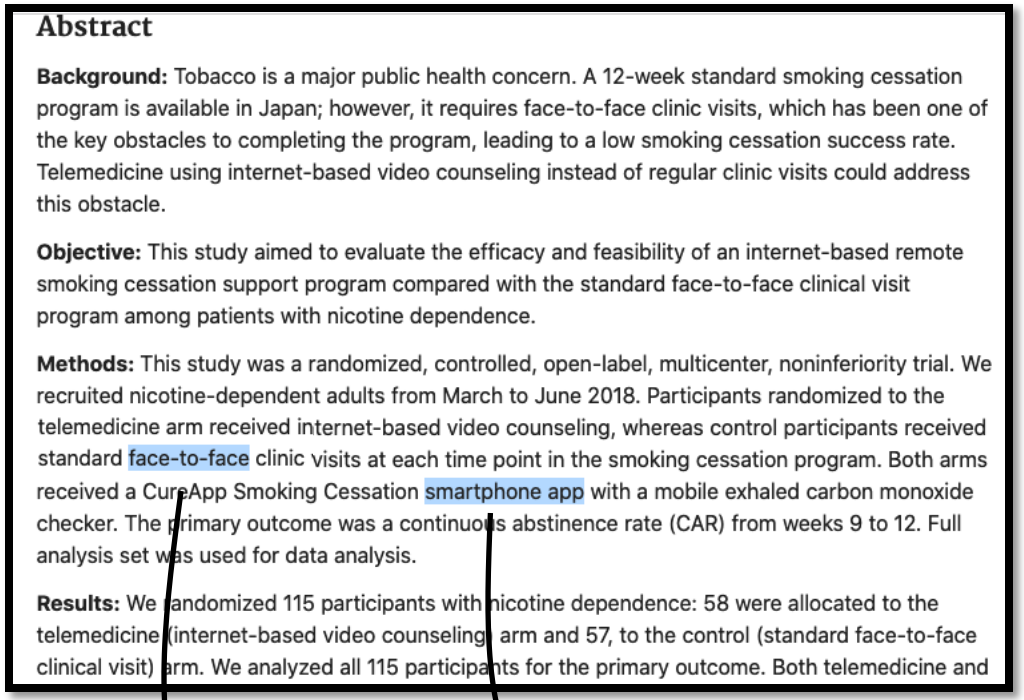
Data: Study Reports



Ontology: Columns and Allowed Values for Data

Universität Zürich[UZH]
Universität St.Gallen School of Medicine

Ontology-based data extraction can have three distinct patterns



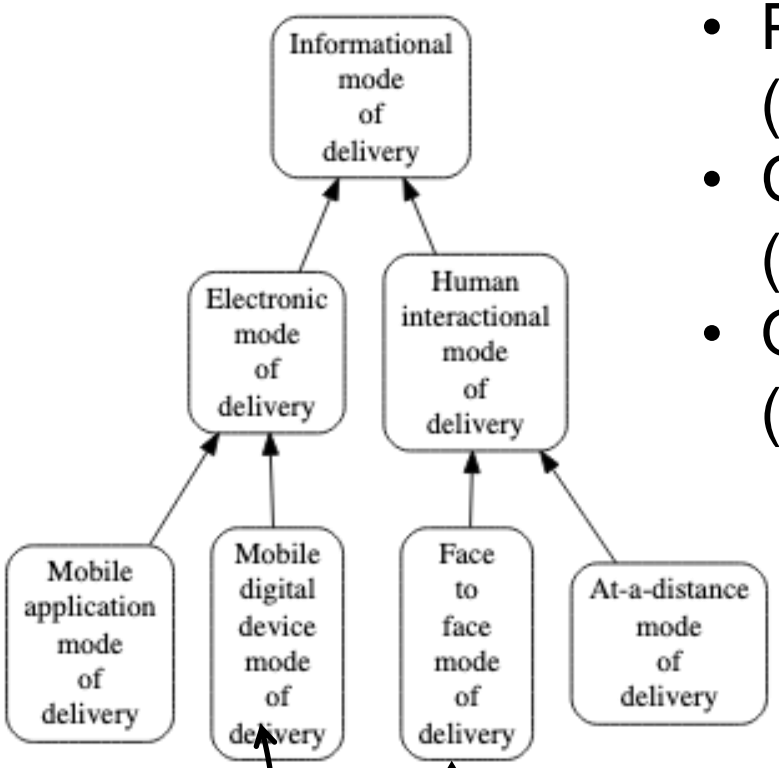
Abstract

Background: Tobacco is a major public health concern. A 12-week standard smoking cessation program is available in Japan; however, it requires face-to-face clinic visits, which has been one of the key obstacles to completing the program, leading to a low smoking cessation success rate. Telemedicine using internet-based video counseling instead of regular clinic visits could address this obstacle.

Objective: This study aimed to evaluate the efficacy and feasibility of an internet-based remote smoking cessation support program compared with the standard face-to-face clinical visit program among patients with nicotine dependence.

Methods: This study was a randomized, controlled, open-label, multicenter, noninferiority trial. We recruited nicotine-dependent adults from March to June 2018. Participants randomized to the telemedicine arm received internet-based video counseling, whereas control participants received standard face-to-face clinic visits at each time point in the smoking cessation program. Both arms received a CureApp Smoking Cessation smartphone app with a mobile exhaled carbon monoxide checker. The primary outcome was a continuous abstinence rate (CAR) from weeks 9 to 12. Full analysis set was used for data analysis.

Results: We randomized 115 participants with nicotine dependence: 58 were allocated to the telemedicine (internet-based video counseling) arm and 57, to the control (standard face-to-face clinical visit) arm. We analyzed all 115 participants for the primary outcome. Both telemedicine and



Annotation types:

- Presence/Absence (e.g. interventions)
- Categorical (e.g. country)
- Quantitative (e.g. mean age)

Can we automate the data extraction part of the review process?

medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

CSH Cold Spring Harbor Laboratory BMJ Yale

HOME | SUBMIT | FAQ | BLOG | ALERTS / RSS | RESOURCES | ABOUT

Search

Advanced Search

Follow this preprint Previous Next

Application of a general LLM-based classification system to retrieve information about oncological trials

Posted December 05, 2024.

Fabio Dennstädt, Paul Windisch, Irina Filchenko, Johannes Zink, Paul Martin Putora, Ahmed Shaheen, Roberto Gaio, Nikola Cihoric, Marie Wosny, Stefanie Aeppli, Max Schmerder, Mohamed Shelan, Janna Hastings

doi: https://doi.org/10.1101/2024.12.03.24318390

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should *not* be used to guide clinical practice.

Download PDF
Print/Save Options
Author Declarations
Supplementary Material
Data/Code
Email
Share
Citation Tools
Get QR code

Dr. Fabio Dennstädt
University of St. Gallen
and Inselspital Bern

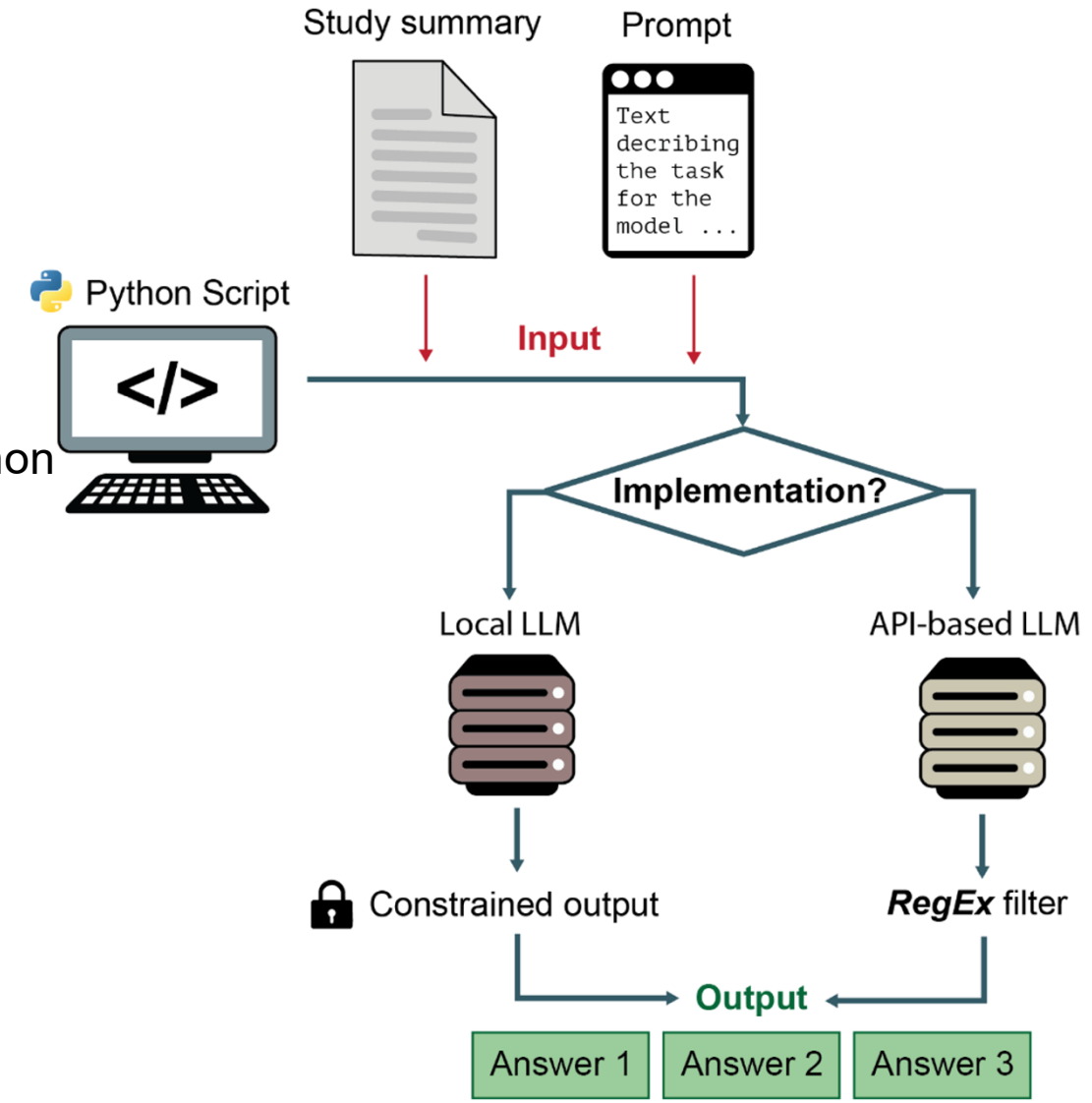
Universität Zürich[UZH]
Universität St.Gallen School of Medicine

Approach: Flexible prompt and constrained output

Flexible text input

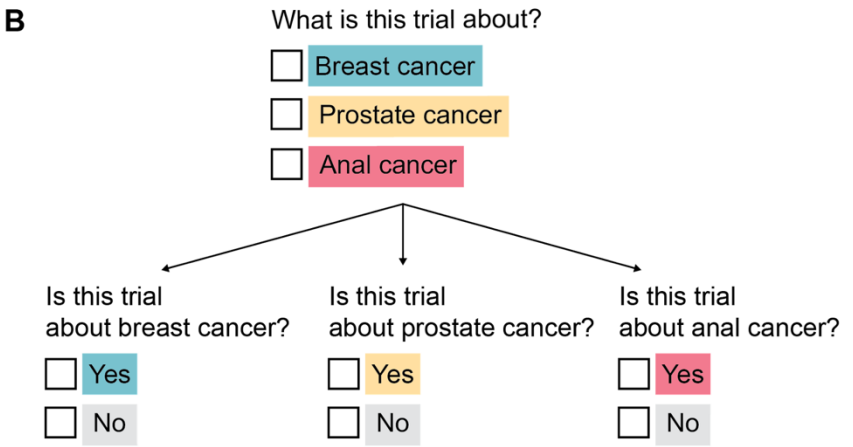
Available as a Python package in PyPI:

general-classifier



Specific classification task with defined allowable values

Decomposed into binary classification tasks



Strategy for evaluation

Open-Source Language Models

for reproducibility and a controlled execution environment

- Mixtral 8x7b-Instruct-0.1
- Meta-Llama-3.1-70B-Instruct
- Qwen2.5-72B-Instruct

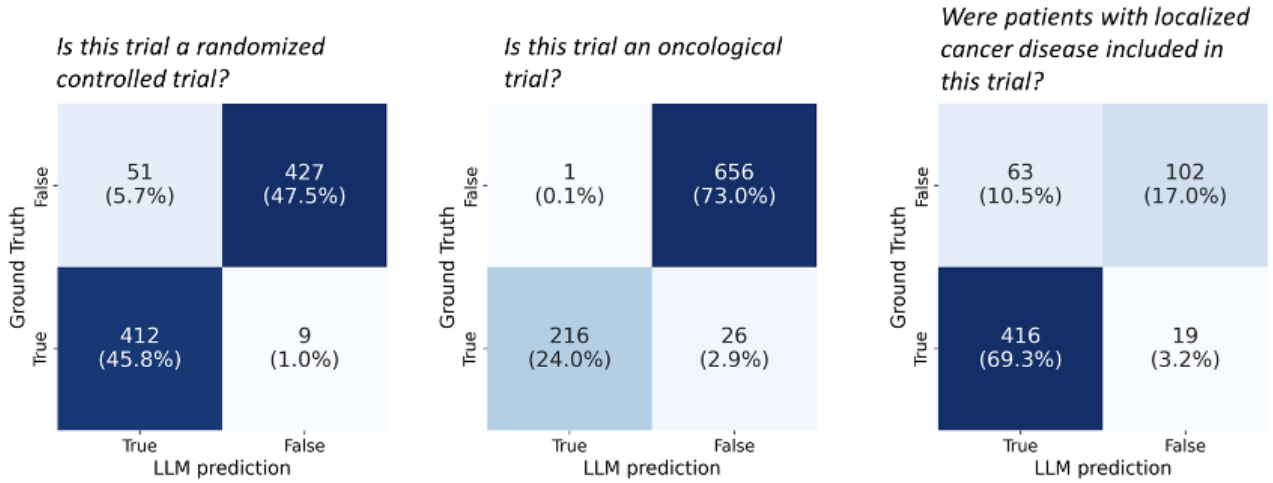
Datasets (from our own previous research)

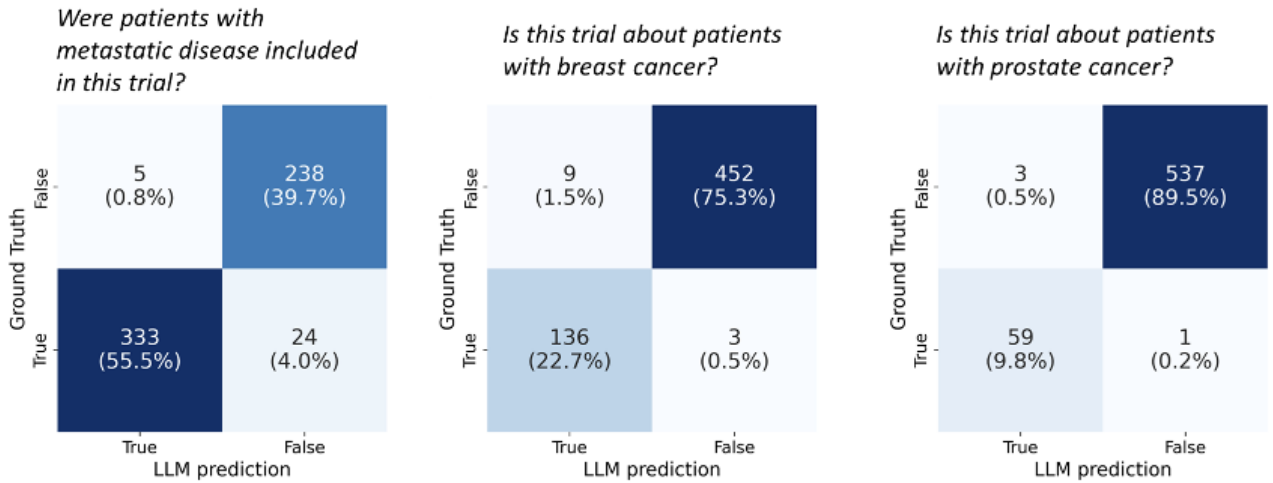
Dataset 1 contains the abstracts of 899 trials with labels whether they are oncological trials and/or RCTs.

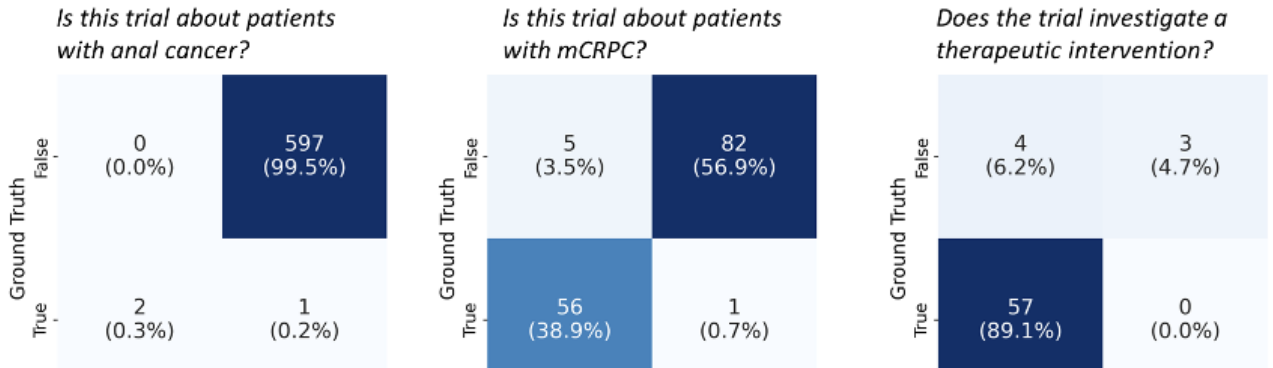
Dataset 2 contains the abstracts of 600 oncological trials with labels whether patients with non-metastatic and/or metastatic disease were included and labels about the type of primary cancer.

Dataset 3 contains abstracts with accompanying information about 144 oncological trials and the label whether it is a trial about patients with mCRPC.

Dataset 4 contains abstracts with accompanying information about 64 oncological trials and the label of whether the trial investigated a therapeutic intervention.







Results

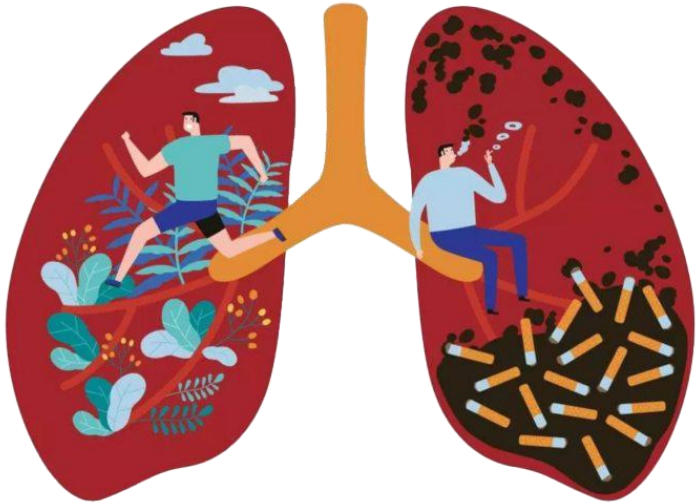
The lowest performance across all models was observed for the question "*Were patients with localized cancer disease included in this trial?*" with accuracies of 86.33% (Mixtral local), 85.49% (Mixtral cloud), 90.50% (Llama3.1), and 77.13% (Qwen2.5).

For the remaining eight questions, all models achieved accuracies above 90%, with the Llama3.1 model reaching accuracies greater than 95%.

Universität Zürich[UZH]

Universität St.Gallen
School of Medicine

Data extraction is hard: – Is tobacco use in the US decreasing?



"In 2019, approximately **20.8%** of U.S. adults (50.6 million) currently used any tobacco product." https://www.cdc.gov/mmwr/volumes/69/wr/mm6946a4.htm

"In 2021, an estimated 46 million (**18.7%**) … U.S. adults currently used any tobacco product" https://www.cdc.gov/tobacco/data-statistics/mmwr/2023-mm7218a1.html

"In 2022, 49.2 million (**19.8%**) … U.S. adults reported current tobacco product use." https://www.cdc.gov/tobacco/php/data-statistics/adult-data-cigarettes/index.html

CDC

- Long term, smoking rates have fallen 73% among adults, from 42.6% in 1965 to 11.6% in 2022.
- Over the last five years, smoking rates have fallen 17% among adults, from 14.0% in 2017.

https://www.lung.org/research/trends-in-lung-disease/tobacco-trends-brief/overall-smoking-trends

American Lung Association

These seem inconsistent, yet they are both accurate. How? → **What is a tobacco product?**

Cox, West, Notley, Soar and Hastings (2023) Addiction 15;118(1):177–188
Towards an ontology of tobacco, nicotine and vaping products

Ontologies enable interpretable machine learning-based prediction

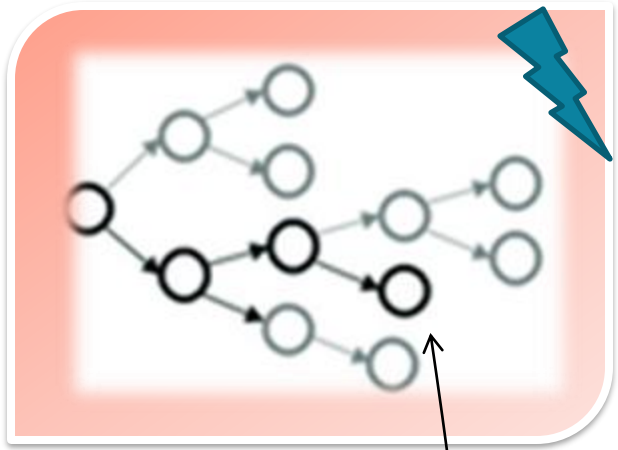
age: 10-30

setting: health care facility

mode of delivery: face to face

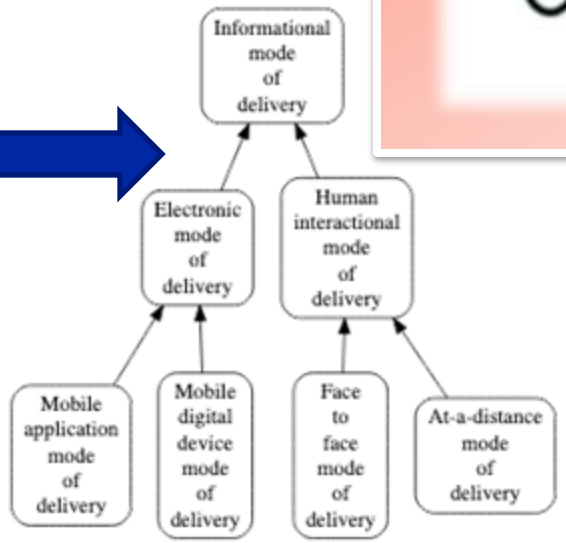
```
Nurse & not Nurse & placebo & not placebo -> 3
```

interpretable machine learning model



predicted outcome

[0.15,0.20]



The trained model enables dynamic prediction of outcomes

Medical Knowledge and Decision Support

Prediction tool part of HBCP

Smoking Cessation

Contact

Intervention
- [ ] 1.1.Goal setting (behavior)
- [ ] 1.2 Problem solving
- [ ] 1.4 Action planning
- [ ] 1.8 Behavioral contract
- [x] 11.1 Pharmacological support
- [ ] 11.2 Reduce negative emotions
- [ ] 12.3 Avoidance/reducing exposure to cues for the behavior
- [ ] 12.5 Adding objects to the environment
- [ ] 12.6 Body changes
- [ ] 13.2 Framing/reframing
- [ ] 2.2 Feedback on behaviour
- [ ] 2.3 Self-monitoring of behavior
- [ ] 2.7 Feedback on outcome(s) of behavior
- [x] 3.1 Social support (unspecified)
- [ ] 3.2 Social support (practical)
- [ ] 4.1 Instruction on how to perform the behavior
- [ ] 4.2 Information about Antecedents
- [ ] 4.5. Advise to change behavior
- [ ] 5.1 Information about health consequences
- [ ] 5.3 Information about social and

Mean age: 15 | 50 | 80 (15 22 29 36 43 50 57 64 71 78 80)

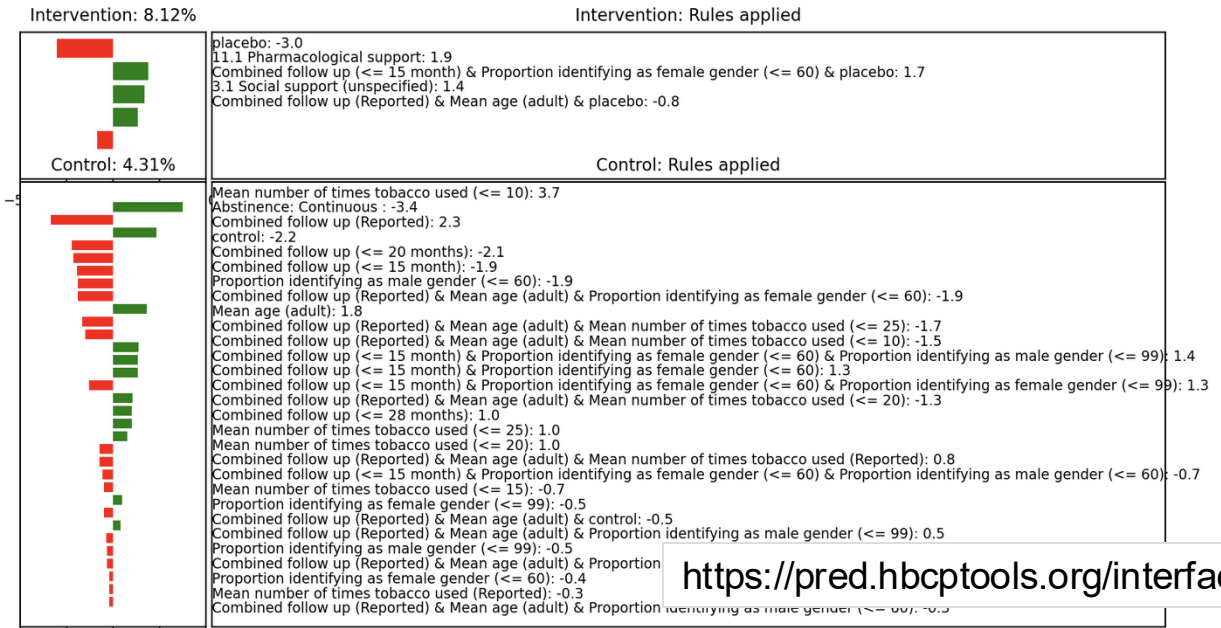
Proportion female: 0 | 50 | 100 (0 10 20 30 40 50 60 70 80 90 100)

Mean number of times tobacco used: 1 | 10 | 30 (1 4 7 10 13 16 19 22 25 28 30)

- [ ] Patient role?

Outcome
- (•) Abstinence: Continuous
- ( ) Abstinence: Point Prevalence
- [ ] Biochemical verification

Follow up (weeks): 4 | 26 | 60 (4 10 16 22 28 34 40 46 52 58 60)



https://pred.hbcptools.org/interface/

Universität St.Gallen School of Medicine

Illustration: Sara Gironi Carnevale (Science Magazine)

Conclusions: Neuro-symbolic approaches to manage medical evidence

- There are many different ways that language models and ontologies can accelerate the management and synthesis of biomedical literature
- However, performance can vary between topics and specific detail-level tasks, thus careful evaluations are key to avoid unsystematic and ad-hoc applications that can degrade the overall quality of the evidence
- There is a need for academic research (not only large commercial organisations) and local initiatives to adapt to local contexts
- Open source models are becoming increasingly competitive and allow secure, reproducible local installations with guardrails built in

Thank you!

Prof. Dr. Janna Hastings

Medical Knowledge and Decision Support

janna.hastings@uzh.ch

@jannahastings.bsky.social

https://hastingslab.org/

Acknowledgements

Marie Wosny
Livia Strasser
Charlotte Tumescheit
Björn Gehrke
Paula Muhr
Fabio Dennstädt
Yating Pan
Maya Braun
Lucas Caluori
Thibault Niederhauser

Martin Glauer
Simon Flügel
Dr Fabian Neuhaus
Prof. Dr. Till Mossakowski
Adel Memariani

Susan Michie
Robert West
James Thomas
Alison Wright
Marta Marques
Paulina Schenk

Wellcome Trust

+ + Many more colleagues and collaborators around the world