Medical Misinformation & Disinformation

Credibility, Expertise, and Prevention

Aleksandra Nabożny, PhD
Polish-Japaneese Academy of Information Technology
NASK National Research Institute

Why This Matters

1. Health misinformation influences treatment decisions
2. Linked to vaccine hesitancy and harmful behaviors
3. Rapid spread via social media
4. Public health outcomes depend on information quality

Ad. 1 & 2

Borges do Nascimento IJ, Pizarro AB, Almeida JM, Azzopardi-Muscat N, Gonçalves MA, Björklund M, Novillo-Ortiz D. Infodemics and health misinformation: a systematic review of reviews. **Bull World Health Organ**. 2022 Sep 1;100(9):544-561. doi: 10.2471/BLT.21.287654. Epub 2022 Jun 30. PMID: 36062247; PMCID: PMC9421549.

Infodemics and health misinformation: a systematic review of reviews

The proportion of health-related misinformation on social media ranged from 0.2% to 28.8%. (Twitter, Facebook, YouTube and Instagram)

The most negative consequences of health misinformation are the

- increase of misleading or incorrect interpretations of available evidence,
- impact on mental health,
- misallocation of health resources
- increase in vaccination hesitancy
- delays in care provision
- increases the occurrence **of hateful and divisive rhetoric**.

Why This Matters

1. Health misinformation influences treatment decisions
2. Linked to vaccine hesitancy and harmful behaviors
3. Rapid spread via social media
4. Public health outcomes depend on information quality

Ad 3

Soroush Vosoughi *et al.*, The spread of true and false news online. *Science***359**,1146-1151(2018).DOI:10.1126/science.aap9559

How fast does misinformation spread?

- 70% higher retweet probability than true news
- Reaches large audiences up to 6× faster
- Produces deeper and broader diffusion cascades
- More likely to go viral
- Driven primarily by human sharing behavior

Agenda:

1. Why truth detection is hard in medicine and what is used instead?
2. How annotation protocols for misinformation are designed?
3. Problems with datasets – inter-rater (dis)agreements
4. Why debunking has limited effectiveness and what to do instead?

Medical Knowledge Is Not Binary

- Probabilistic and uncertain
- Context-dependent
- Evolving over time
- Example: changing COVID-19 guidelines

Credibility Instead of Truth

- Truth often unavailable or contested
- Use proxies: source authority, evidence use, transparency, alignment with scientific consensus

CREDIBILITY

"the fact that someone or something can be believed or trusted"

TRUTH

"the real facts about a situation, event, or person"

Humans Confuse Facts and Opinions

Pew Research Center, June, 2018, "Distinguishing Between Factual and Opinion Statements in the News"

Study on 5000 American citizents analyzes the ability to classify 12 statements as facts (e.g., "2 + 2 = 4") or opinions (e.g., "Green is the most beautiful color"), with an emphasis on political examples.

- approximately **46% of participants did not perform better than chance**,
- the errors were driven by "partisan bias" — statements favoring one's own political group were more likely to be classified as facts.

Humans Confuse Facts and Opinions

Leonardo Bursztyn, Aakaash Rao, Christopher Roth, David Yanagizawa-Drott, Opinions as Facts, The Review of Economic Studies, Volume 90, Issue 4, July 2023

- The experiment shows that individuals with different ideological views choose opinion programs rather than “straight news,” even when incentivized to learn factual information
- context: the early stage of the COVID-19 pandemic in the United States.
- **Clear differences** were found in the **adoption of preventive behaviors** among viewers of the two most popular opinion programs aired on the same network, which presented conflicting narratives about the threat.
- It was also shown that **areas with relatively higher viewership of programs downplaying the risk experienced higher numbers of COVID-19 cases** and deaths.

Humans Confuse Facts and Opinions

My studies: expert assesments of text samples; ‚non-credible' labels marked as red, ‚credible' samples marked as green

"As Dr. Malcolm Kendrick argues, stress and social isolation are often overlooked sources of heart disease (…)"

"According to Prof. Kuna, patients with asthma should also stop using nebulizers, because in the case of infection with the SARS-CoV-2 virus, an aerosol will be generated that will spread throughout the entire room."

"(...) the ratio of the inactive to active form of OC (UCR) was significantly elevated in individuals using statins, indicating vitamin K deficiency. **According to the researchers**, statins also affected the international normalized ratio (INR) and interacted with vitamin K antagonists."

Framing Effect

- Classic lung-cancer treatment study (McNeil et al., 1982)

➡️ Participants preferred surgery much more under the **survival frame**
➡️ Participants preferred radiation more under the **mortality frame**

Even clinicians are influenced by wording

Agenda:

1. Why truth detection is hard in medicine and what is used instead?
2. How annotation protocols for misinformation are designed?
3. Problems with datasets – inter-rater (dis)agreements
4. Why debunking has limited effectiveness and what to do instead?

Tagging manipulation and intention

Based on: Arkadiusz Modzelewski, Giovanni Da San Martino, Pavel Savov, Magdalena Anna Wilczyńska, and Adam Wierzbicki. 2024. MIPD: Exploring Manipulation and Intention In a Novel Corpus of Polish Disinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19769–19785, Miami, Florida, USA. Association for Computational Linguistics.

Cherry Picking (CHP) Presenting information utilizing only data that supports a given hypothesis or argument, while Ignoring the broader context

Quote Mining (QM) Using a short fragment of someone's longer speech in a way that significantly distorts its original tone

Anecdote (AN) The use of evidence in the form of personal experience or an isolated case, possibly rumor or hearsay, most often to discredit statistics

Tagging manipulation and intention

Based on: Arkadiusz Modzelewski, Giovanni Da San Martino, Pavel Savov, Magdalena Anna Wilczyńska, and Adam Wierzbicki. 2024. MIPD: Exploring Manipulation and Intention In a Novel Corpus of Polish Disinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19769–19785, Miami, Florida, USA. Association for Computational Linguistics.

Whataboutism (WH). Responding to a substantive argument not by addressing the heart of the matter, but by raising a new point that is unrelated

to the topic at hand

Misleading Clickbait (MC). A technique involves giving a title to the text that misrepresents or contradicts the content discussed within the article. Title created with a purpose to attract attention

Tagging manipulation and intention

Based on: Arkadiusz Modzelewski, Giovanni Da San Martino, Pavel Savov, Magdalena Anna Wilczyńska, and Adam Wierzbicki. 2024. MIPD: Exploring Manipulation and Intention In a Novel Corpus of Polish Disinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19769–19785, Miami, Florida, USA. Association for Computational Linguistics.

Appeal to Emotion (AE). The use of words and phrases that are to arouse in the recipient extreme emotion and attitude to the presented matter

False Cause (FC). The individual employing this technique assumes a cause-and-effect relationship solely based on the observed correlation

Exaggeration (EG). The author overstates a phenomenon, making it appear larger, better, or worse, or oversimplifies a phenomenon making it seem less significant or smaller than it truly is

Tagging manipulation and intention

Based on: Arkadiusz Modzelewski, Giovanni Da San Martino, Pavel Savov, Magdalena Anna Wilczyńska, and Adam Wierzbicki. 2024. MIPD: Exploring Manipulation and Intention In a Novel Corpus of Polish Disinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19769–19785, Miami, Florida, USA. Association for Computational Linguistics.

Reference Error (RE). In this technique, the author refers to fake experts, propaganda statements made by politicians, anonymous entries published on social media, or false quotes from famous people to authenticate the presented thesis It may present false choices and false analogies.

Strawman (ST) It involves distorting someone else's argument in a way that makes it easier to refute it. It is often done by attributing a stance to opponents, who do not share it

Leading Questions (LQ) Flooding the recipient with a series of consecutive suggestive questions or putting them together leads the recipient to a predetermined thesis

What about the Online Health Content (OHI)

- Misinterpreting scientific uncertainty
- „BigPharma” and other conspirational narratives
- Hedging – strategy absent in general domain
- Advertisements!

Data Is the Bottleneck for AI

- Machine learning requires labeled datasets
- Annotation is expensive and slow
- Requires domain expertise

Agenda:

1. Why truth detection is hard in medicine and what is used instead?
2. How annotation protocols for misinformation are designed?
3. Problems with datasets – inter-rater (dis)agreements
4. Why debunking has limited effectiveness and what to do instead?

No Standard Label Scheme

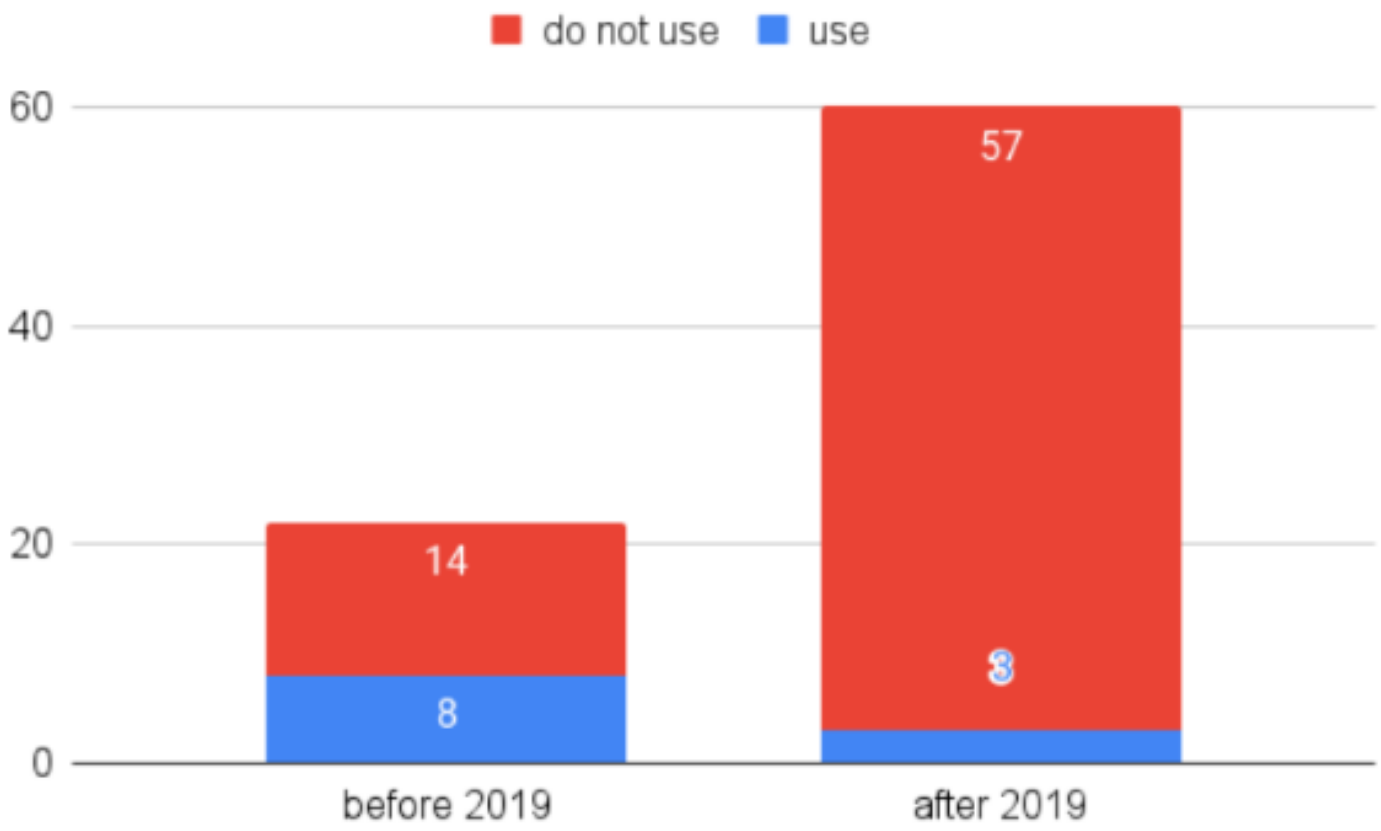
- There are medically-approved quality assesment protocols such as DISCERN, EQIP or HealthNewsReviews.org but they are hardly used in Computer Science studies

D. Charnock, S. Shepperd, G. Needham, and R. Gann. 1999. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. Journal of Epidemiology & Community Health 53, 2 (2 1999), 105–111. doi:10.1136/jech.53.2.105

A.I. Charvet-Berard, P. Chopard, and T.V. Perneger. 2008. Measuring quality of patient information documents with an expanded EQIP scale. Patient Education and Counseling 70, 3 (3 2008), 407–411. doi:10.1016/j.pec.2007.11.018

Gary Schwitzer. 2008. How Do US Journalists Cover Treatments, Tests, Products, and Procedures? An Evaluation of 500 Stories. PLoS Medicine 5, 5 (5 2008), e95. doi:10.1371/journal.pmed.0050095

No Standard Label Scheme



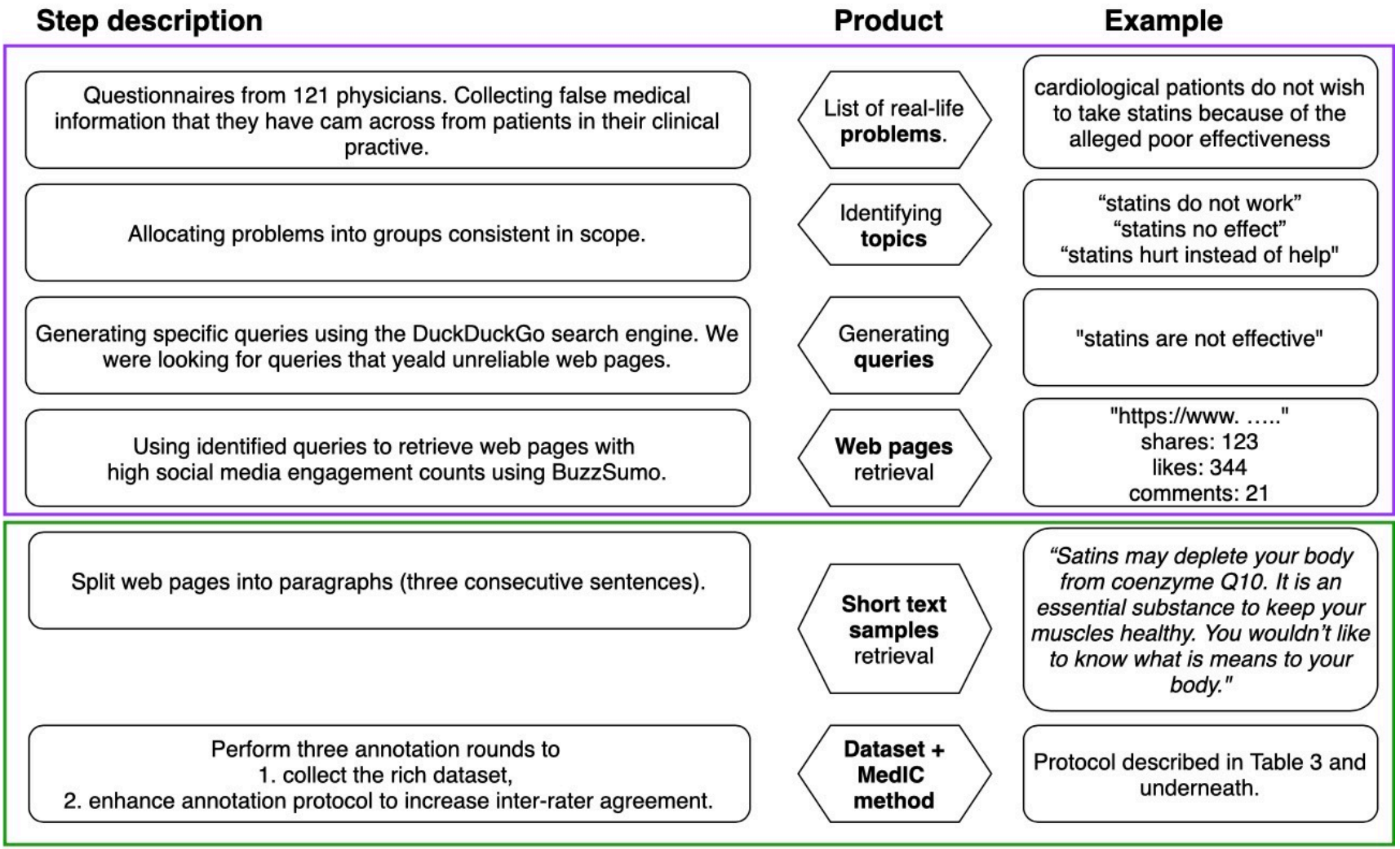
Limited Expert Collaboration

- Many datasets use crowd workers or students
- Experts are costly and scarce
- Complex guidelines needed

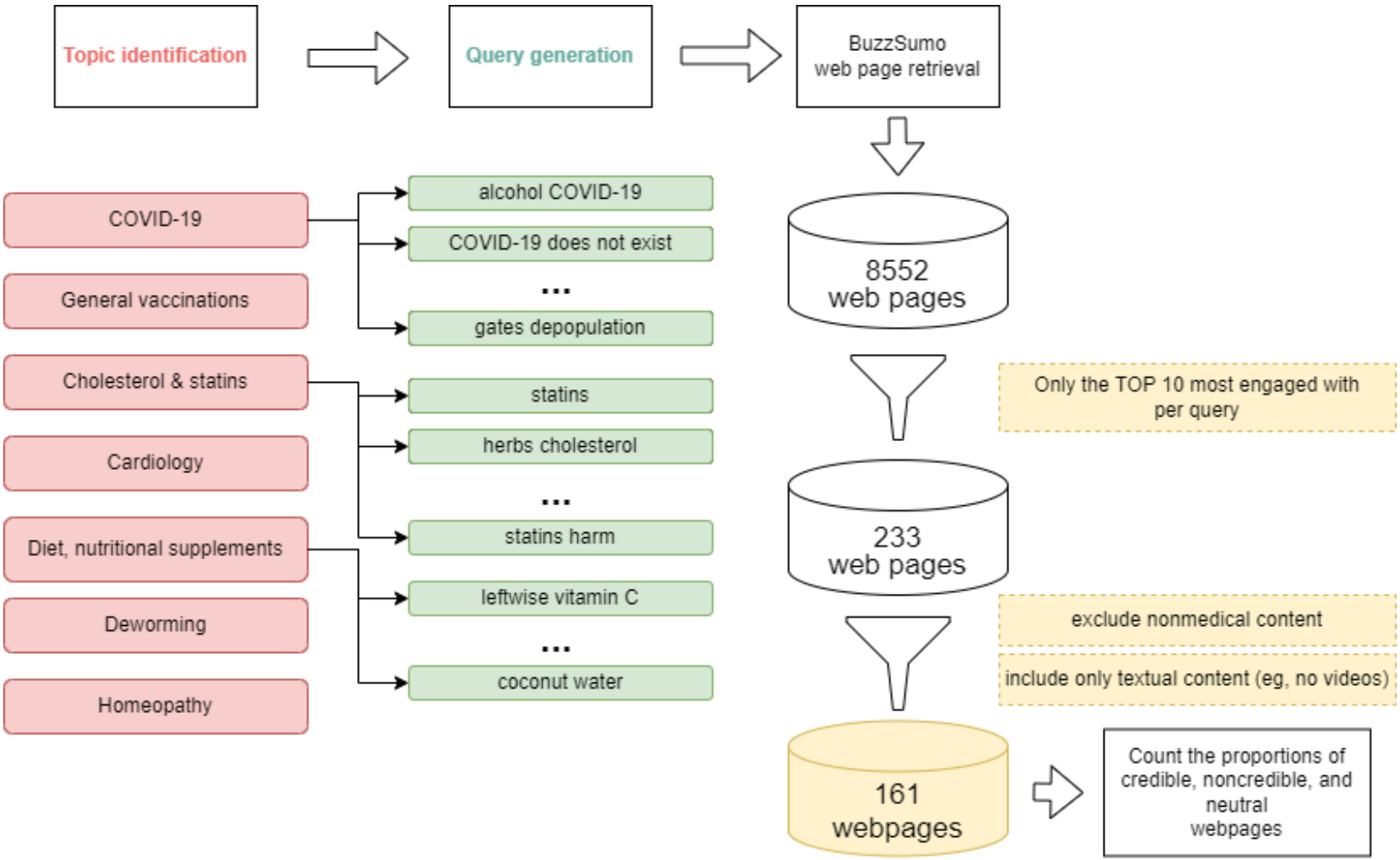
… experts are used, and…

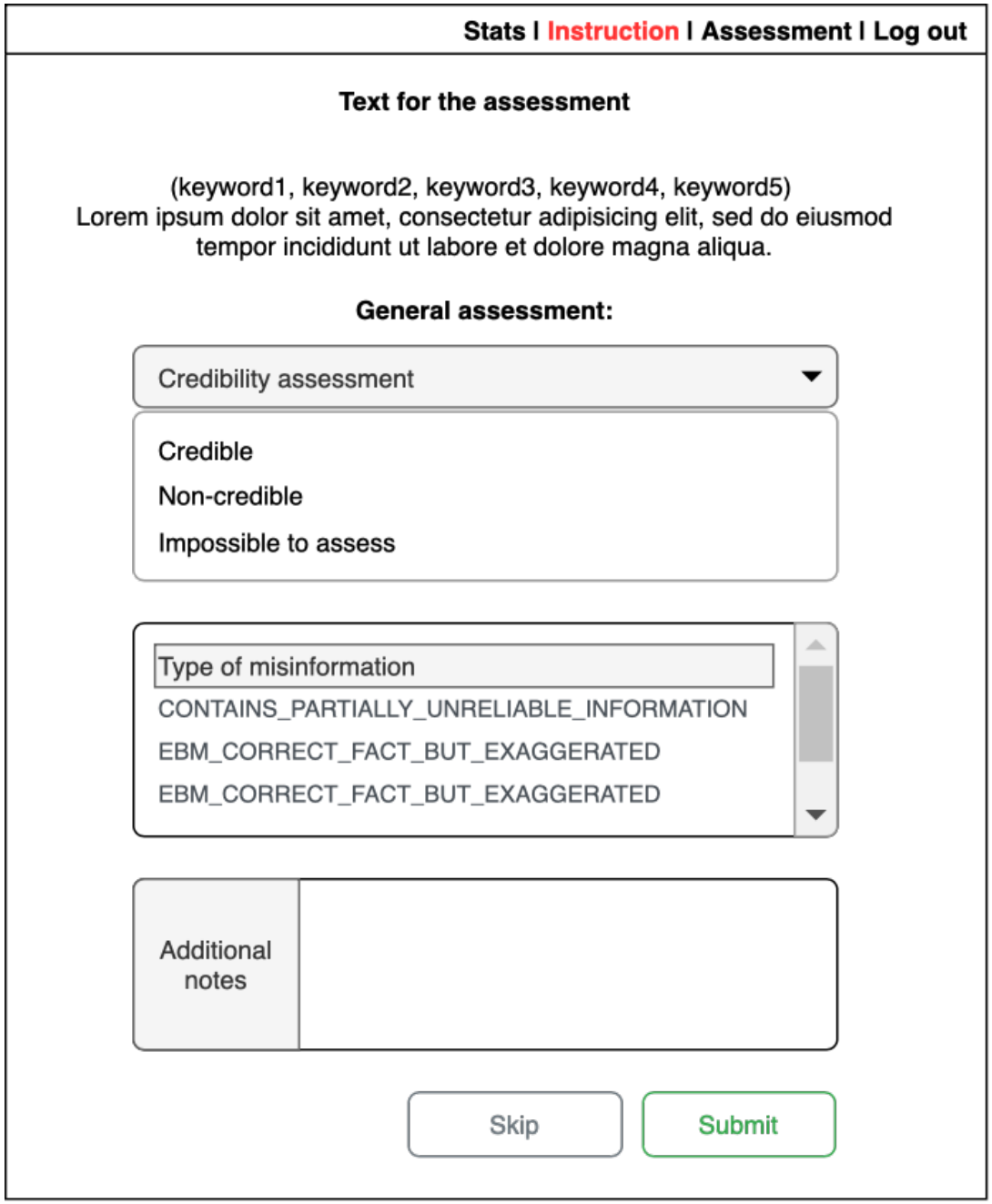
- Often small samples
- Limited reporting of agreement
- Subjectivity remains

Experts Do Not Always Agree

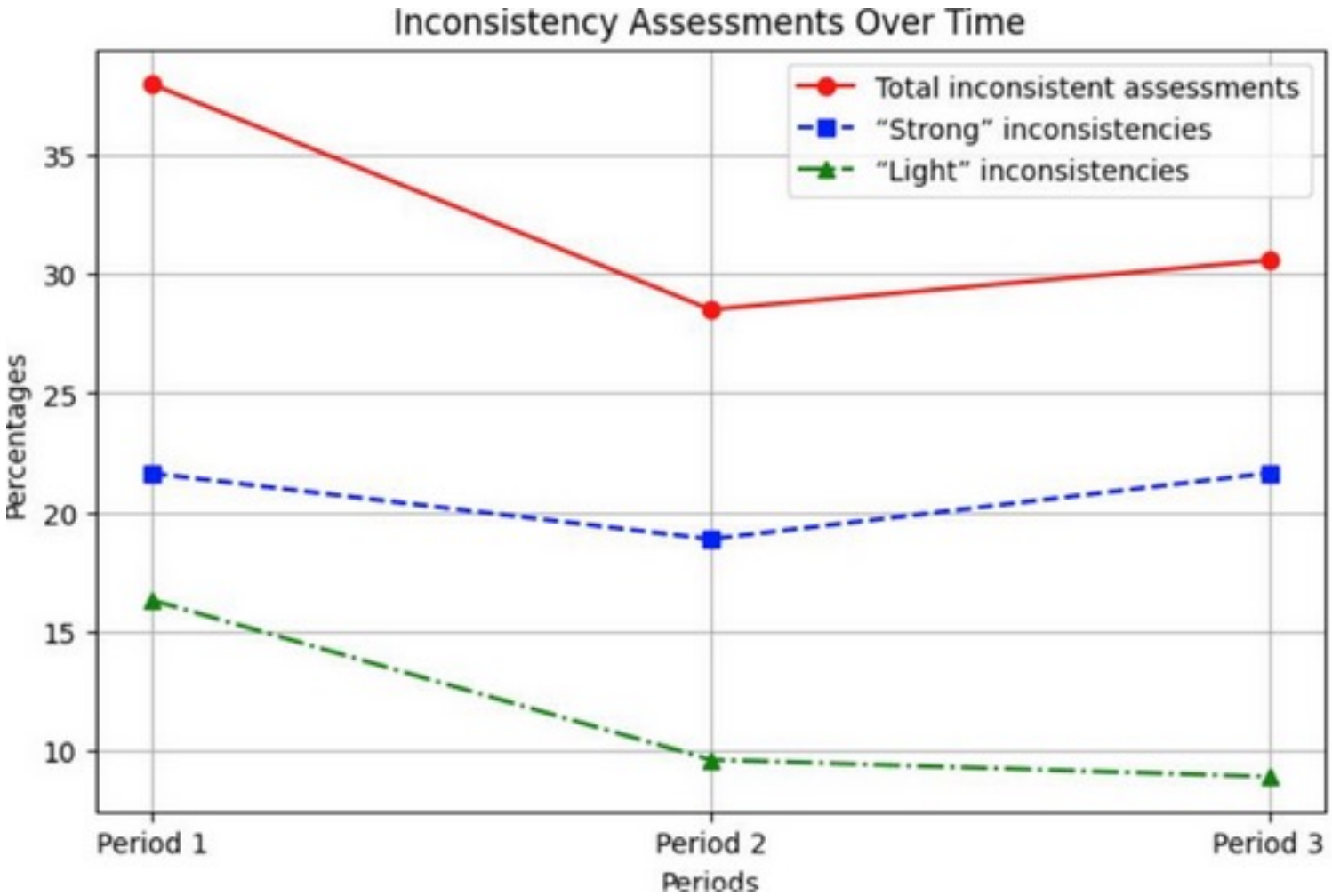


Experts Do Not Always Agree




Stats I Instruction I Assessment I Log out
Text for the assessment
(keyword1, keyword2, keyword3, keyword4, keyword5)
Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.
General assessment:
Credibility assessment
Credible
Non-credible
Impossible to assess
Type of misinformation
CONTAINS_PARTIALLY_UNRELIABLE_INFORMATION
EBM_CORRECT_FACT_BUT_EXAGGERATED
EBM_CORRECT_FACT_BUT_EXAGGERATED
Additional notes
Skip
Submit

Experts Do Not Always Agree



Sample analysis

Sample:
“Only when diet and other treatment methods prove ineffective is it recommended to introduce medications that affect lipid metabolism, and their dosage is individually determined by a physician for each patient. Statins (hydroxymethylglutaryl-CoA reductase inhibitors) block the HMG-CoA reductase enzyme, which plays an important role in cholesterol synthesis. They lower LDL cholesterol and total cholesterol levels and cause a slight increase in HDL levels.”

Sample analysis

Sample:
“**Only when diet and other treatment methods prove ineffective** is it recommended to introduce medications that affect lipid metabolism, and their dosage is individually determined by a physician for each patient. Statins (hydroxymethylglutaryl-CoA reductase inhibitors) block the HMG-CoA reductase enzyme, which plays an important role in cholesterol synthesis. They lower LDL cholesterol and total cholesterol levels and cause a slight increase in HDL levels.”

Expert post-commentary

“Most of the content is reliable. The discrepancy arises from the statement, ‘Only when diet and other treatments prove ineffective is it recommended to introduce medications that affect lipid metabolism.’ According to guidelines for the prevention of cardiovascular diseases, in certain clinical situations physicians must prescribe statins immediately, without waiting for improvement from diet.”

„Disagreement Evidence" for medicine

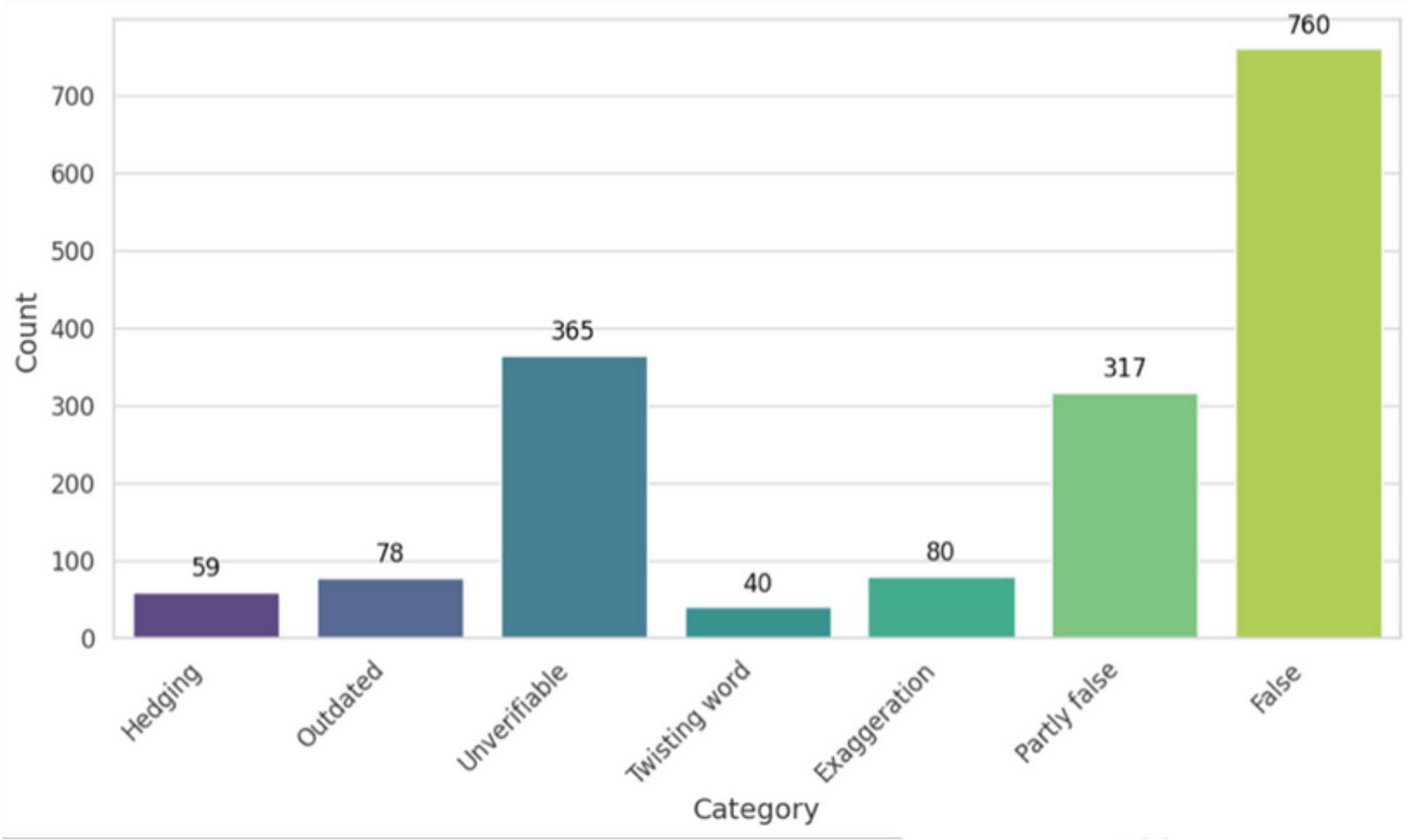
- Hatzoglou et al. (2016): retrospective analysis of initial and second-opinion radiology reports for 300 neuroimaging studies, **19% disagreement**
- Fechtenbaum et al. (2017): inter-expert agreement in vertebral fracture diagnosis, **30% disagreement**

these tasks involve well-defined imaging protocols!

Why Experts Disagree

- Different experience and training
- Risk tolerance differences
- Rhetorical framing that is hard to interpret

Why Experts Disagree



Implication for AI

- Ground truth may be uncertain
- Model performance is fundamentally limited

Agenda:

1. Why truth detection is hard in medicine and what is used instead?
2. How annotation protocols for misinformation are designed?
3. Problems with datasets – inter-rater (dis)agreements
4. Why debunking has limited effectiveness and what to do instead?

Debunking vs Prebunking

- Debunking: correction after exposure
- Prebunking: build resistance beforehand

Inoculation Theory

- Expose to weakened misinformation
- Build psychological immunity

The Bad News Game

The Bad News Game

- Players simulate misinformation creators
- Learn manipulation techniques firsthand

Roozenbeek, J., & van der Linden, S. (2019). **Fake news game confers psychological resistance against online misinformation.** *Palgrave Communications*, 5, Article 65. https://doi.org/10.1057/s41599-019-0279-9

Evidence of Effectiveness

- Improves recognition of misinformation techniques
- Reduces perceived reliability of false content
- Roozenbeek & van der Linden (2019)

Limits of Prebunking

- Effects may decay over time
- Not a standalone solution
- Best combined with education and policy
- **Best combined with early detection** when the game could be updated with „fresh”, emergent topics

Conclusions

- Need standardized datasets
- Interdisciplinary collaboration
- Transparent reporting

Implications for Policy Makers

- Media literacy education
- Platform accountability
- Public health campaigns

Key Takeaways

- Medical misinformation is not just false information
- Involves uncertainty, cognition, communication, and trust
- Credibility is best assessed by assesing source credibility and the **intention**

Discussion Questions

- Can AI reliably detect harmful medical claims?
- Is prevention more realistic than detection?
- How should platforms respond?

Thank you!

Aleksandra Nabożny, PhD
NASK National Research Institute
aleksandra.nabozny@gmail.com

DISCORD CHANNEL:
https://discord.gg/yWfrB7cz