# Capturing Semantic Gaps in MeSH through Human-AI Collaboration



**Jian Qin** **Bei Yu** **Qiaoyi Liu**

School of Information Studies

Syracuse University

United States

NKOS Workshop at DC2026, Seoul, South Korea, August 3, 2024


# Talking points

- MeSH and indexing of medical journal articles
- The case of indexing public health policy and COVID-19 vaccination publications
- Observations and discussion
  - Controlled vocabulary vs. author keywords
  - Biomedical phenomena and clinical interventions vs. social/political debates
  - Automatic indexing vs. vocabulary updated-ness
- Final thoughts

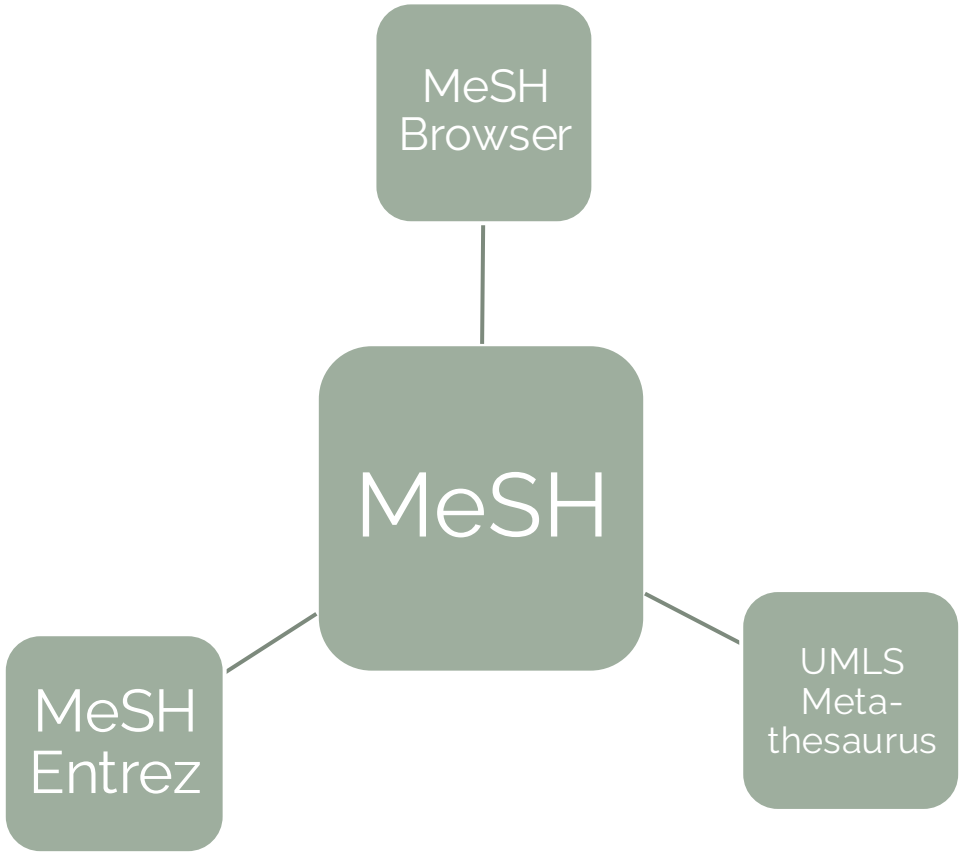
# MeSH and indexing of medical journal articles

MTIX (Medical Text Indexer-NeXt Generation)

MTIA (Medical Text Indexer-Automated)

# MeSH and use in indexing and search

- A controlled vocabulary for indexing, cataloging, and searching for biomedical and health-related information and documents.
- Various online systems provide access to MeSH and the vocabulary is available in several NLM online systems.
- Focus on biomedical phenomena and clinical interventions



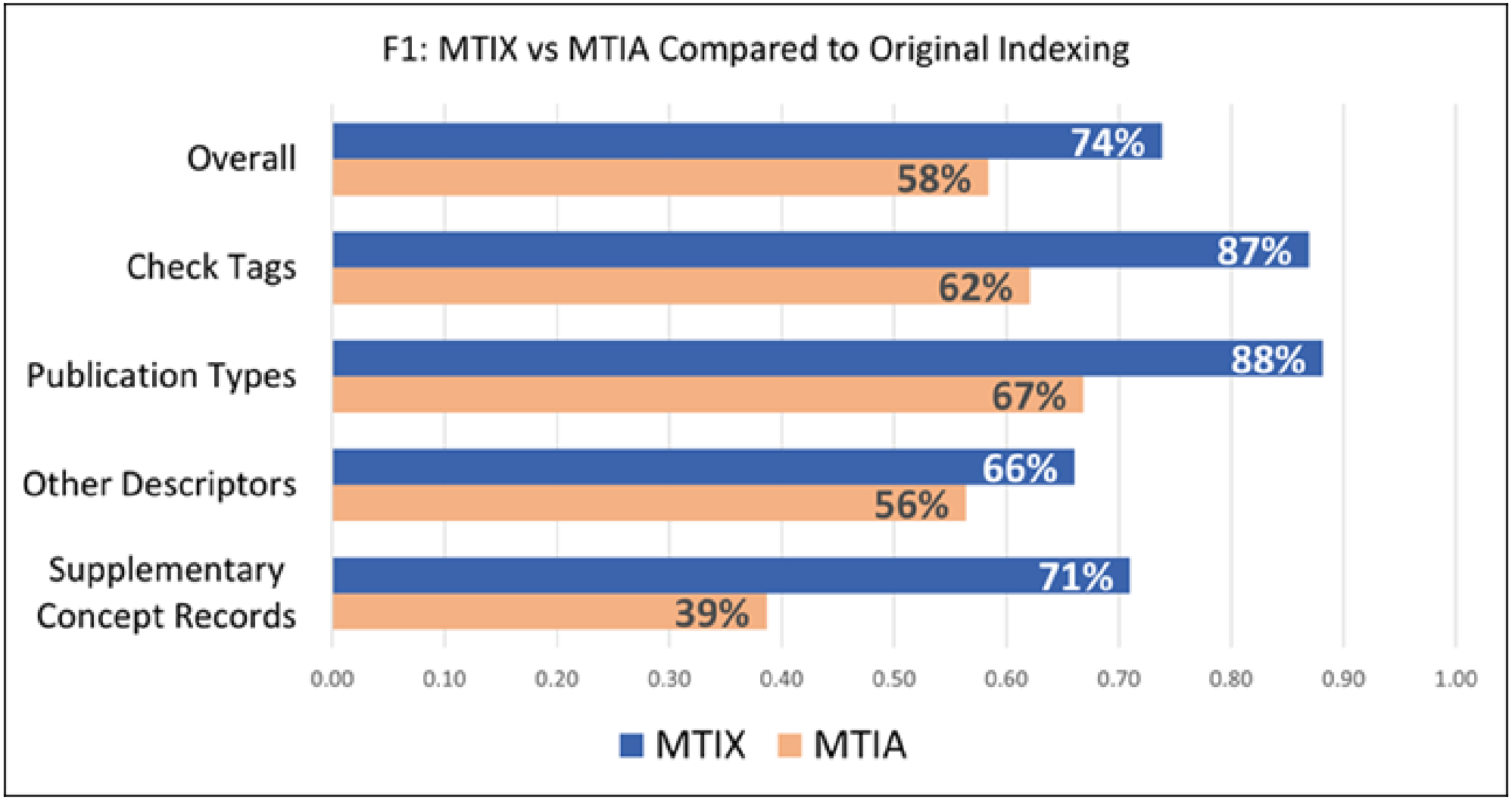
# Automatic indexing and updates of MeSH

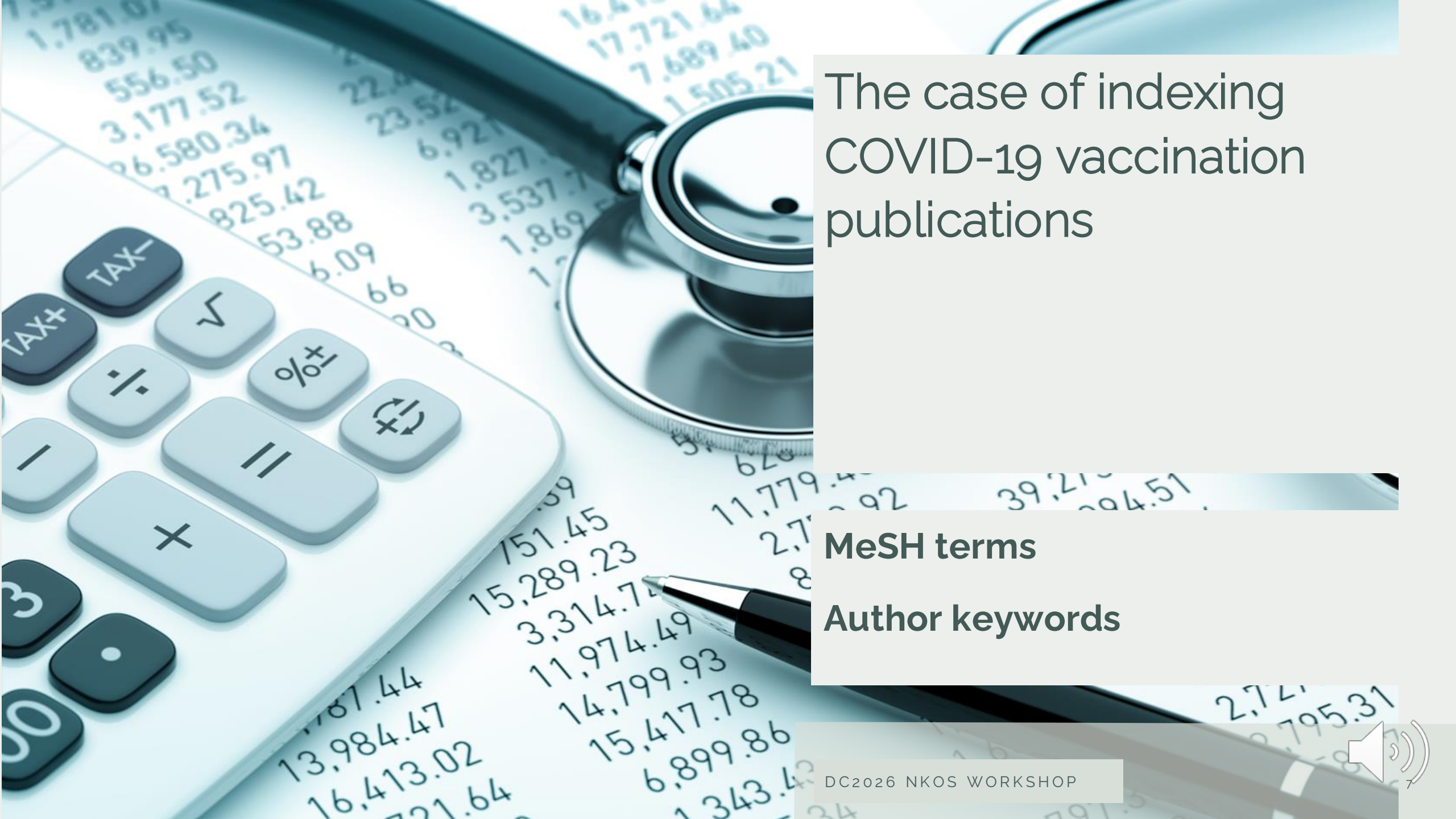
- MTIA (Medical Text Indexer-Automated) algorithm
  - Originated in 1996 and expanded during the 2010s
  - Transitioned to full automation in 2022

**Rule-based**

- MTIX (M edical Text Indexer-NeXt Generation) algorithm
  - An AI system to automatically index MEDLINE citations
  - Introduced in Aprile 2024, which replaced MTIA

**Neural network-based**


F1: MTIX vs MTIA Compared to Original Indexing
Overall
74%
58%
Check Tags
87%
62%
Publication Types
88%
67%
Other Descriptors
66%
56%
Supplementary Concept Records
71%
39%
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
MTIX
MTIA



# The case of indexing COVID-19 vaccination publications

**MeSH terms**

**Author keywords**

**Topic:** public health policy and Covid-19 vaccines

**Query:** "(public health policies) AND (COVID-19 Vaccines[MeSH Terms])" retrieved: 3186 articles

- No author keywords: 1075 (33.7%) articles
- HAVE Both MeSH terms and Author KWS: 2111 (66.26%)

- Number of unique MeSH terms:
  - 4336 before normalization, 1753 after normalization
- Number of unique Author KWs:
  - 4769 before normalization, 4066 after normalization

**Example of MeSH term normalization:**

**Before:** *COVID-19/economics/epidemiology/prevention & control

**After:** covid-19, economics, epidemiology, prevention & control

- Further normalization used "pubmedbert-base-embeddings" model to match semantically similar terms

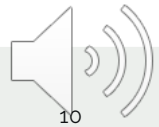
# Finding 1: MeSH and KW overlaps

| Term Category | Frequency | Percent |
|---|---|---|
| Concept | 434 | 60.20% |
| Geo | 100 | 13.87% |
| Disease | 97 | 13.45% |
| Human | 48 | 6.66% |
| Vaccine | 42 | 5.83% |
| **Total** | **721** | **100%** |

- Co-occurred terms in both normalized MeSH and KW: 495
  - Increased to 721 when setting the similarity threshold at 0.90
- KW “substance use disorder” and MeSH term “substance-related disorder” is 0.80
  - A 0.80 threshold increase co-occurred terms to 1349, which accounts for 77% of MeSH terms and 33% KWs

# Finding 2: Structural mismatch

- COVID-19 vaccine debate is fundamentally a debate about *attitudes, trust, rights, and equity*, but MeSH was built to describe *biomedical phenomena and clinical interventions*.
- The example of risk-related terms
  - KWs: 34 risk-related, e.g., risk perception, risk communication, benefit-risk balance, risk acceptance
  - MeSH: one descriptor—Risk and subheadings



# MeSH Tree Structure for descriptor Risk

Investigative Techniques [E05]
- Epidemiologic Methods [E05.318]
  - Statistics as Topic [E05.318.740]
    - Probability [E05.318.740.600]
      - Bayes Theorem [E05.318.740.600.200]
      - Likelihood Functions [E05.318.740.600.400]
      - Markov Chains [E05.318.740.600.500]
      - Odds Ratio [E05.318.740.600.600]
      - Position-Specific Scoring Matrices [E05.318.740.600.650]
      - Propensity Score [E05.318.740.600.675]
      - Proportional Hazards Models [E05.318.740.600.700]
      - **Risk [E05.318.740.600.800]**
        - Logistic Models [E05.318.740.600.800.450]
        - Protective Factors [E05.318.740.600.800.582]
        - Risk Assessment [E05.318.740.600.800.715]
        - Risk Factors [E05.318.740.600.800.725]
      - Uncertainty [E05.318.740.600.900]

Mathematical Concepts [G17]
- Probability [G17.680]
  - Odds Ratio [G17.680.600]
  - **Risk [G17.680.7**
  - Uncertainty [G17.680.900]

Health Care Quality, Access, and Evaluation [N05]
- Quality of Health Care [N05.715]
  - Health Care Evaluation Mechanisms [N05.715.360]
    - Statistics as Topic [N05.715.360.750]
      - Probability [N05.715.360.750.625]
        - Bayes Theorem [N05.715.360.750.625.150]
        - Likelihood Functions [N05.715.360.750.625.450]
        - Markov Chains [N05.715.360.750.625.500]
        - Odds Ratio [N05.715.360.750.625.590]
        - Propensity Score [N05.715.360.750.625.620]
        - Proportional Hazards Models [N05.715.360.750.625.650]
        - **Risk [N05.715.360.750.625.700]**
          - Logistic Models [N05.715.360.750.625.700.450]
          - Protective Factors [N05.715.360.750.625.700.570]
          - Risk Assessment [N05.715.360.750.625.700.690]
          - Risk Factors [N05.715.360.750.625.700.700]
        - Uncertainty [N05.715.360.750.625.850]

Environment and Public Health [N06]
- Public Health [N06.850]
  - Epidemiologic Methods [N06.850.520]
    - Statistics as Topic [N06.850.520.830]
      - Probability [N06.850.520.830.600]
        - Bayes Theorem [N06.850.520.830.600.200]
        - Likelihood Functions [N06.850.520.830.600.400]
        - Markov Chains [N06.850.520.830.600.500]
        - Odds Ratio [N06.850.520.830.600.600]
        - Propensity Score [N06.850.520.830.600.650]
        - Proportional Hazards Models [N06.850.520.830.600.700]
        - **Risk [N06.850.520.830.600.800]**
          - Logistic Models [N06.850.520.830.600.800.450]
          - Protective Factors [N06.850.520.830.600.800.582]
          - Risk Assessment [N06.850.520.830.600.800.715]
          - Risk Factors [N06.850.520.830.600.800.725]
        - Uncertainty [N06.850.520.830.600.900]

## KWs containing Risk

| | | |
|---|---|---|
| Acceptable risk | Perceived risk of infection | Risk groups |
| Assessment, Risk benefit | Populations at risk | Risk management |
| Behavioral risk | Resurgence risk | Risk mitigation |
| Benefit-risk | Risk acceptance | Risk of death |
| Benefit-risk balance | Risk attitudes | Risk perception |
| Crisis and emergency risk communication | Risk awareness | Risk perceptions |
| Epidemiological risk | Risk capacity | Risk reduction |
| Health risk | Risk communication | Risk taking |
| Health risk behavior | Risk compensation | Social risk factors |
| High-risk groups | Risk factor | Vaccine risk |
| High-risk patients | | |

# Finding 3: Conceptual gaps

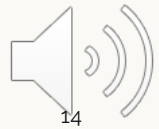
- MeSH dominates on structural/administrative categories
  - Demographic descriptors, study design labels
- KWs dominate on the conceptual frontier
  - Specific policy instruments (“mandates”, “passports”)
  - Info-ecosystem concepts (“misinformation”, “conspiracy theories”)
  - Communication strategies (“risk communication”, “vaccination campaign”)


COVID-19
VIRUS
VACCINE
20 ml
LOT
COVID-19
VIRUS
VACCINE
20 ml
LOT
COVID-19
VIRUS
VACCINE
20 ml
LOT
BASE
BASE

# Discussion

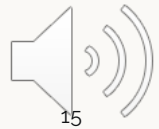
# The Classical Conundrum: Controlled Vocabulary vs. Author Keywords

- Pros and cons of controlled vocabulary vs. keywords in information retrieval have extensive research in the last sixty years
- Findings from this study did not add anything new in this regard, but show that KWs can supply much more enriched content representation for the social, political, and behavioral aspects of the vaccine debate, which MeSH falls short
- The MTIX algorithm indexing results have room for improvement, MeSH and other knowledge organization systems will be more important in fully automatic indexing systems

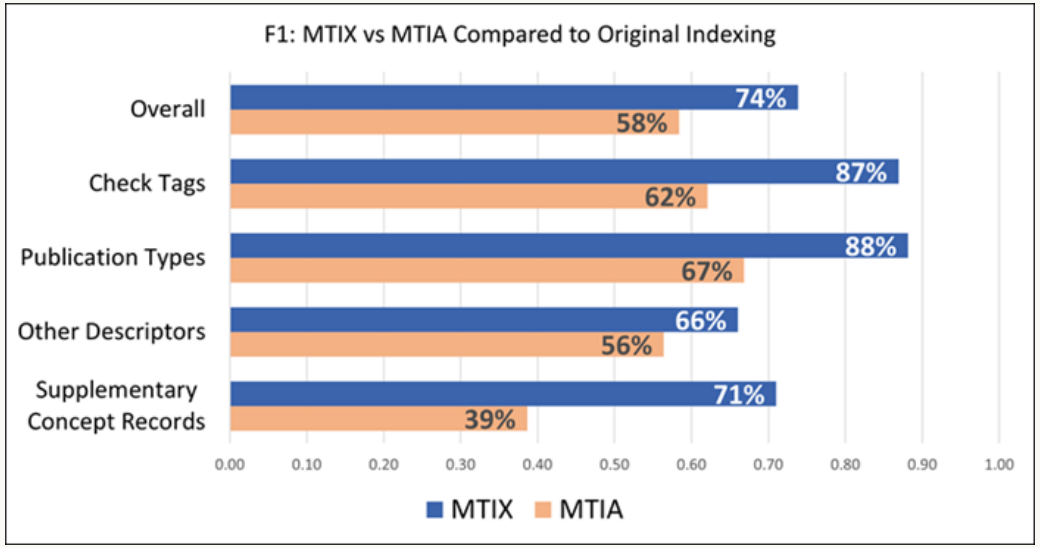


# Biomedical Phenomena and Clinical Interventions vs. Social/Political/Behavioral Debates

- MeSH's focus is on biomedical phenomena and clinical interventions
- Only general terms are used for social/behavioral/policy aspects of medical/health topics
- Should it expand its vocabulary coverage for socially/behaviorally significant, medical/health related topics?
- Will KWs be sufficient for retrieving literature on such topics?



# Automatic indexing vs. vocabulary coverage




- Automatic indexing with MTIX algorithm heavily relies on MeSH to map synonyms, related terms, etc.
- MeSH is being vigorously updated each year
- General terms for social/behavioral aspects of biomedical and clinical topics can retain decent recall but not precision
- KWs play a significant role in retrieving precision where MeSH falls short

# CONCLUSION

- The empirical findings on COVID-19 vaccines and public health policy show only a small portion of indexing overlapping between MeSH and KWs
- Where MeSH offers only broad descriptors for the vaccine debate, KWs provide richer, more specific terms
- Large Language Models are a powerful tool for studying chunks of MeSH or other KOS for targeted subject area updates

# Thank you!

# Question?

**Jian Qin, Bei Yu, Qiaoyi Liu**
School of Information Studies
Syracuse University

Syracuse, New York, United States