Semantic Citation: Mentions of Knowledge in Citation Contexts

Tzu-Kun Hsiao

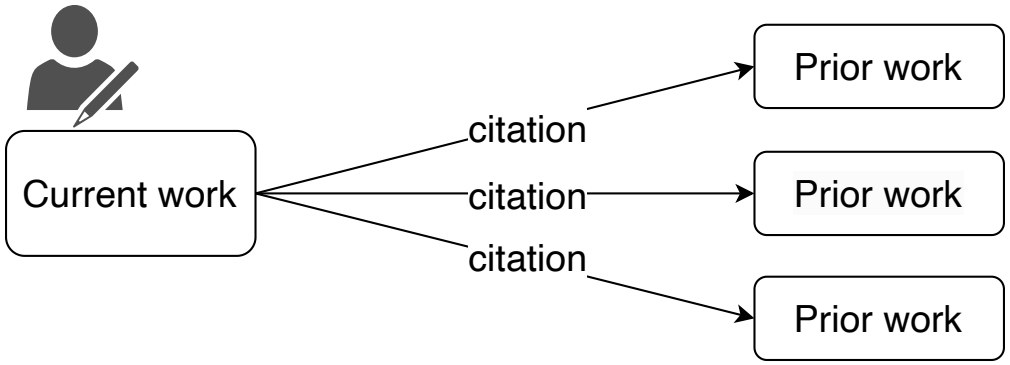
Committee:
Associate Professor Vetle I. Torvik, School of Information Sciences, Chair
Associate Professor Jodi Schneider, School of Information Sciences
Associate Professor Halil Kilicoglu, School of Information Sciences
Assistant Professor Matthew Turk, School of Information Sciences & Astronomy

Outline

- Motivation
- Accomplished work
- Proposed work
- Timeline

Why should we study citation contexts?

- Citations are commonly regarded as intellectual linkages between research articles (Smith, 1981).



- A substantial number of citation studies were based on bibliographic records.
- Citations may have different meanings for their citing authors.
- Citation context analysis can help us understand citations better.

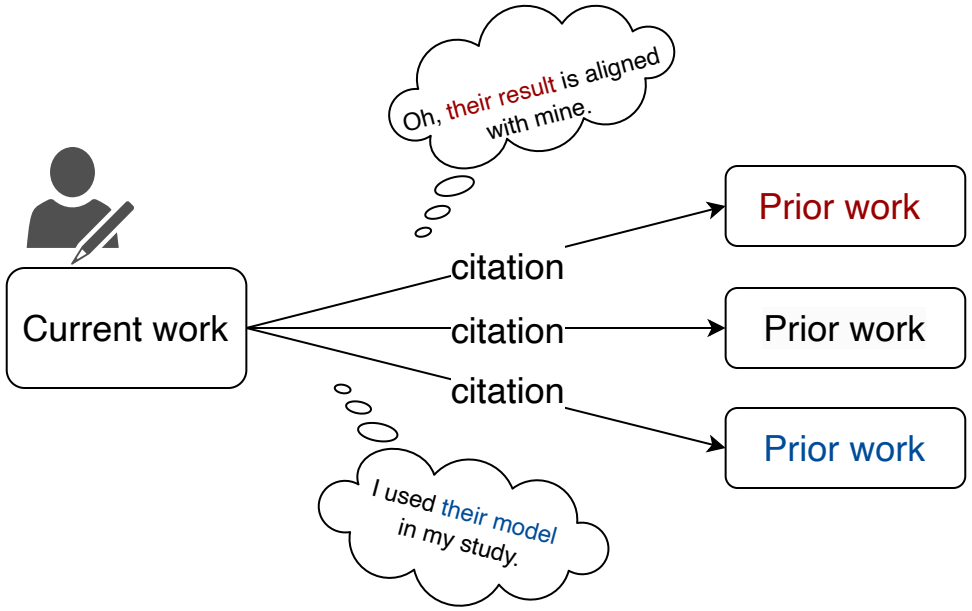
Why should we study citation contexts?

- Citations are commonly regarded as intellectual linkages between research articles (Smith, 1981).

- A substantial number of citation studies were based on bibliographic records.
- Citations may have different meanings for their citing authors.
- Citation context analysis can help us understand citations better.

Why should we study citation contexts?

- Citations are commonly regarded as intellectual linkages between research articles (Smith, 1981).



- A substantial number of citation studies were based on bibliographic records.
- Citations may have different meanings for their citing authors.
- Citation context analysis can help us understand citations better.

Citation context datasets

- A citation context dataset: a dataset containing sentences in academic articles with inline citations identified.
- A citation context dataset can be constructed from:
  - PDF files
    - Science Parse (Cohan et al., 2019), ParsCit (Jurgens et al., 2018), and GROBID (Khadka & Knoth, 2018; Lo et al., 2020).
    - Failure to expand citation ranges
      (e.g., missing citation [2] and [3] in citations "[1-4]")
  - LaTeX files and XML files
    - Parsers developed by the datasets' authors
- Lack of clear documentation of parsers' performance
- Datasets may be used as-is in future research (Wu et al., 2022).
- Difficulties in updating or customizing the datasets

Motivation

- Address the challenges of obtaining high-quality citation context data.
  - Produce a high-quality, large-scale dataset.
  - A clear documentation of the pipeline
  - A quality evaluation

Motivation

- Address the challenges of obtaining high-quality citation context data.
  - Produce a high-quality, large-scale dataset.
  - A clear documentation of the pipeline
  - A quality evaluation

- **RQ 1:**
  What makes it difficult to produce a citation context dataset from the Journal Article Tag Suite (JATS) XML files created by a wide variety of content providers (e.g., publishers)?
  - Why JATS?
  - Nuances of JATS XML files

Motivation

- Study how existing knowledge has been mentioned in scientific literature.
  - Problematic science
  - Diversity of citation contexts mentioning the same work

Motivation

- Study how existing knowledge has been mentioned in scientific literature.
  - Problematic science
  - Diversity of citation contexts mentioning the same work

- **RQ 2:**
  How can citation contexts contribute to the advancement of citation analysis?
  - RQ 2a: Why do retracted articles get cited after their retractions?
  - RQ 2b: How diverse are the citation contexts when an article receives multiple citations?

RQ 1 - Accomplished work: OpCitance

Hsiao, T.-K., & Torvik, V. I. (2023). OpCitance: Citation contexts identified from the PubMed Central Open Access articles. *Scientific Data*, *10*, 243. https://doi.org/10.1038/s41597-023-02134-x

- Built from PubMed Central (PMC) open access (OA) articles
- Address the challenges of handling nuances in JATS XML files.
  - The pipeline is available to the public.
- A quality evaluation confirms the dataset's soundness.

OpCitance: Challenges addressed

OpCitance: Challenges addressed

- Nested structure and interchangeable use of references' IDs

PMCID:5952554

1. For reviews on isocyanates and their use in industry see: , For polyurethane application see:
   (a) Six C. and Richter F., *Ullmann's Encyclopedia of Industrial Chemistry*, Wiley-VCH, 2012, vol. 20, pp. 63–82. [Google Scholar]
   (b) *Isocyanates, Organic. Kirk-Othmer Encyclopedia of Chemical Technology*, Wiley, New York, 3rd edn, 1982, vol. 19, pp. 28–62. [Google Scholar]
   (c) Wang Z., *Comprehensive Organic Name Reactions and Reagents*, Wiley, New York, 2009, pp.1772–1774. [Google Scholar]
   (d) Engels H.-W., Pirkl H.-G., Albers R., Albach R. W., Krause J., Hoffman A., Casselmann H., Dormish J. *Angew. Chem., Int. Ed.* 2013;52:9422. [PubMed] [Google Scholar]

```
▼<ref-list>
 ▼<ref id="cit1">
  ▶<note>
   ...
   </note>
  ▶<mixed-citation publication-type="journal" id="cit1a">
   ...
   </mixed-citation>
  ▶<mixed-citation publication-type="journal" id="cit1b">
   ...
   </mixed-citation>
  ▶<mixed-citation publication-type="book" id="cit1c">
   ...
   </mixed-citation>
  ▶<element-citation publication-type="journal" id="cit1d">
   ...
   </element-citation>
  </ref>
```

Since their discovery in 1848 by Wurtz, isocyanates have received significant attention from the synthetic community.1 Isocyanates are important bulk and fine chemicals and are used industrially in coatings, paints, foams, adhesives, elastomers, and as building blocks for pharmaceuticals and agrochemicals. Over 4000 isocyanates are also commercially available. Perhaps most notably, isocyanates are used to form polyurethanes. In 2011, 14 million tons of polyurethanes were produced, corresponding to *ca.* 5% of the global polymer market.1*d* Consequently, it is difficult to downplay the industrial importance of isocyanates.1*d*

```
Since their discovery in 1848 by Wurtz, isocyanates have received significant
attention from the synthetic community.
<xref rid="cit1" ref-type="bibr">1</xref>
Isocyanates are important bulk and fine chemicals and are used industrially in
coatings, paints, foams, adhesives, elastomers, and as building blocks for
pharmaceuticals and agrochemicals. Over 4000 isocyanates are also commercially
available. Perhaps most notably, isocyanates are used to form polyurethanes. In
2011, 14 million tons of polyurethanes were produced, corresponding to
<italic>ca.</italic>
5% of the global polymer market.
<xref rid="cit1d" ref-type="bibr">
 1
 <italic>d</italic>
</xref>
Consequently, it is difficult to downplay the industrial importance of isocyanates.
```

OpCitance: Challenges addressed

- Nested structure and interchangeable use of references' IDs
- The implicitly mentioned citations

PMCID:4067518

The theory that drastically decreased pregnancy rates as a direct result of the modern woman's lifestyle, including the availability of contraceptives, contributes to an increased incidence of endometriosis is well established [9-11].

The "hyphen" could be:

- a hyphen
- an en dash (Unicode character U + 2013)
- a minus sign (Unicode character U+ 2212)
- two hyphens/en dashes/minus signs.

OpCitance: Challenges addressed

- Nested structure and interchangeable use of references' IDs
- The implicitly mentioned citations
- Customized code for sentence tokenization

NLTK can break a sentence containing abbreviations into fragmnets:

> *The clinical presentation of scar endometriosis, i.e.*

> *, tender swellings, mimics other dermatological and/or surgical conditions and delays the diagnosis.*

After correction:

> *The clinical presentation of scar endometriosis, i.e., tender swellings, mimics other dermatological and/or surgical conditions and delays the diagnosis.*

What's in OpCitance?

- All sentences in 2,049,871 articles
- 137 million inline citations are annotated.
- Locations of citation contexts are provided.
- Order of the sentences are preserved.
  - Citation context window size can be customized.

pmcid	pmid	location	IMRaD	sentence_id	total_sentences	intxt_id	intxt_pmid	intxt_pmid_source	intxt_mark	best_id	best_source	best_id_diff	citation	progression
5219817	28090276	body	I	1	56	-	-	-	-	-	-	-	Maintenance of calcium/phosphate equilibrium and bone resorption regulation in human body are very significant processes that affected by different parameters.	1.79
5219817	28090276	body	I	2	56	5219817_B1	-	-	1>|B1|	-	-	NONE	Parathyroid hormone consisted of 84 amino acids and secreted by parathyroid glands which is among the main mechanisms of the body that involve in regulation of such a complex process |B1|.	3.57
5219817	28090276	body	I	3	56	-	-	-	-	-	-	-	Bone regeneration process takes 3 to 6 months and involves the coupling of bone resorption and formation pathways.	5.36
5219817	28090276	body	I	4	56	5219817_B1	-	-	1>|B1|	-	-	NONE	If for any reason this process fails, gradually the person will suffer from osteoporosis |B1|, |B2|.	7.14
5219817	28090276	body	I	4	56	5219817_B2	21111078	xml,pmc	2>|B2|	21111078	xml,ice,pmc,len,pat,dim	SAME	If for any reason this process fails, gradually the person will suffer from osteoporosis |B1|, |B2|.	7.14

What's in OpCitance?

- All sentences in 2,049,871 articles
- 137 million inline citations are annotated.
- Locations of citation contexts are provided.
- Order of the sentences are preserved.
  - Citation context window size can be customized.

pmcid	pmid	location	IMRaD	sentence_id	total_sentences	intxt_id	intxt_pmid	intxt_pmid_source	intxt_mark	best_id	best_source	best_id_diff	citation	progression
5219817	28090276	body	I	1	56	-	-	-	-	-	-	-	Maintenance of calcium/phosphate equilibrium and bone resorption regulation in human body are very significant processes that affected by different parameters.	1.79
5219817	28090276	body	I	2	56	5219817_B1	-	-	1>|B1|	-	-	NONE	Parathyroid hormone consisted of 84 amino acids and secreted by parathyroid glands which is among the main mechanisms of the body that involve in regulation of such a complex process |B1|.	3.57
5219817	28090276	body	I	3	56	-	-	-	-	-	-	-	Bone regeneration process takes 3 to 6 months and involves the coupling of bone resorption and formation pathways.	5.36
5219817	28090276	body	I	4	56	5219817_B1	-	-	1>|B1|	-	-	NONE	If for any reason this process fails, gradually the person will suffer from osteoporosis |B1|, |B2|.	7.14
5219817	28090276	body	I	4	56	5219817_B2	21111078	xml,pmc	2>|B2|	21111078	xml,ice,pmc,len,pat,dim	SAME	If for any reason this process fails, gradually the person will suffer from osteoporosis |B1|, |B2|.	7.14

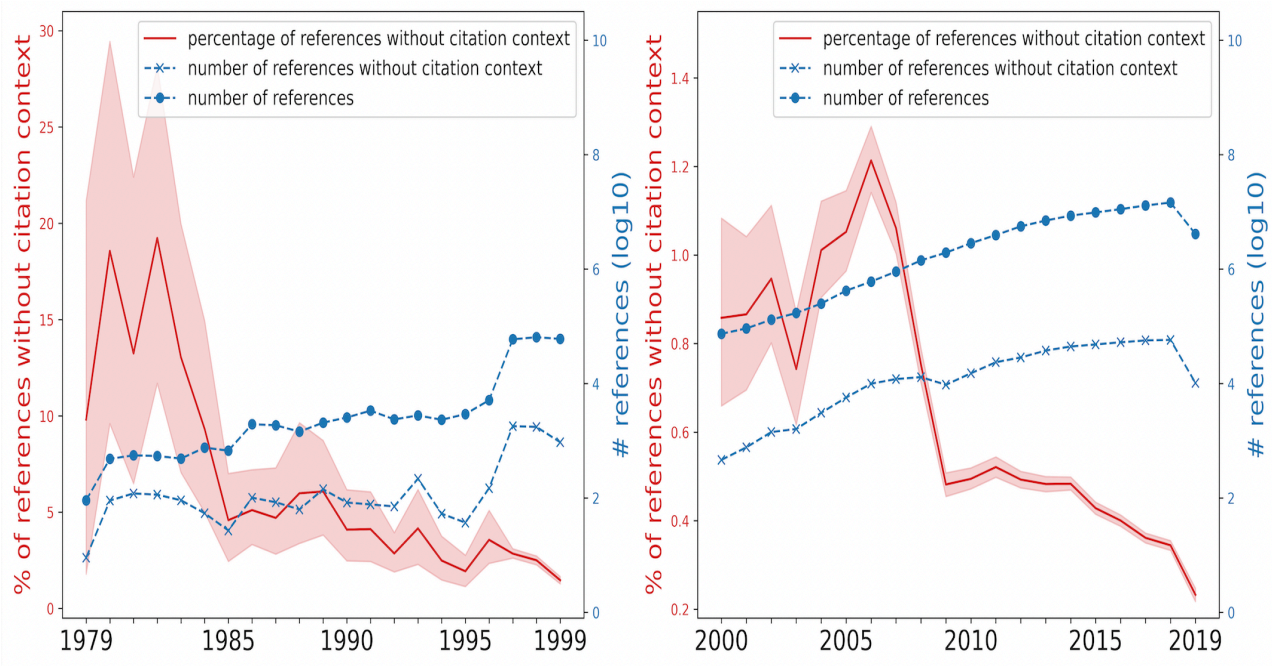
OpCitance: Quality evaluation

- Inline citation identification rate: 99.49%
- Factors that may be related to unidentified inline citations

OpCitance: Quality evaluation

- Inline citation identification rate: 99.49%
- Factors that may be related to unidentified inline citations

- Year of publication



OpCitance: Quality evaluation

- Inline citation identification rate: 99.49%
- Factors that may be related to unidentified inline citations

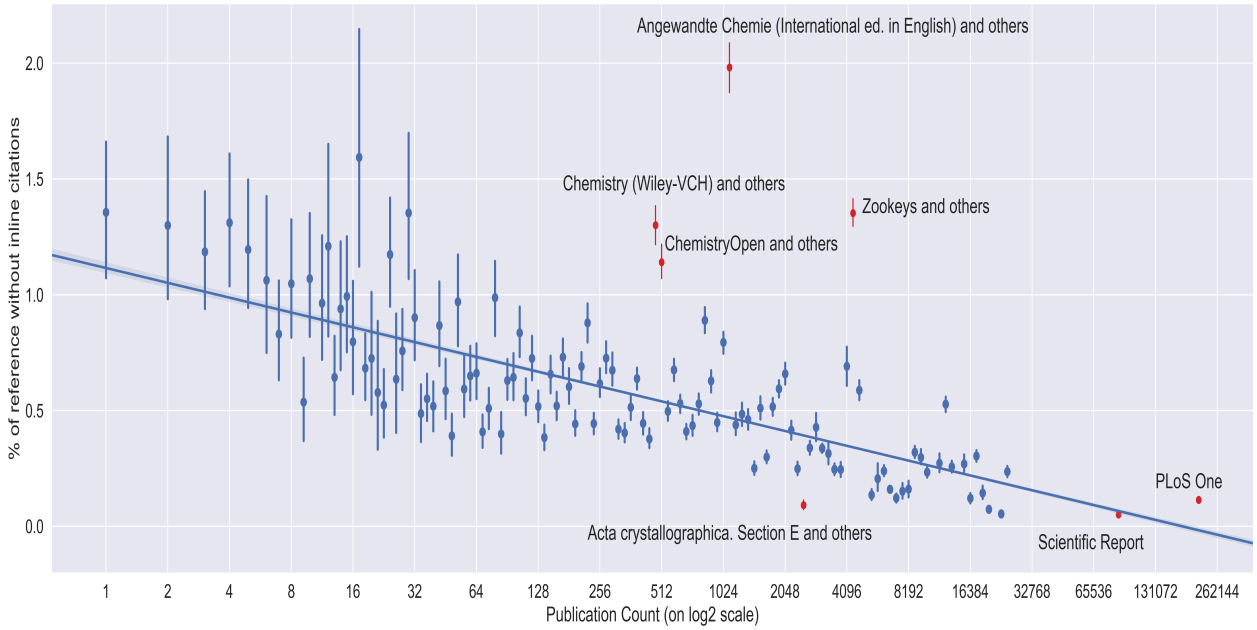
- Year of publication
- Deposit source

Deposit Source	# Articles	# References	# References without inline citations (%)
Full participation	1,830,722	76,275,089	325,347 (0.43)
Selective deposit	138,356	6,763,865	87,587 (1.29)
NIH portfolio	58,357	2,395,980	17,018 (0.71)
Author manuscript	22,436	1,038,412	7,519 (0.72)
Total	2,049,871	86,473,346	437,471 (0.51)

OpCitance: Quality evaluation

- Inline citation identification rate: 99.49%
- Factors that may be related to unidentified inline citations

- Year of publication
- Deposit source
- Journal size



RQ 2a: Citations to retracted articles

RQ 2a: Why do retracted articles get cited after their retractions?

- Accomplished work

  Hsiao, T.-K., & Schneider, J. (2021). Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine. *Quantitative Science Studies, 2*(4), 1144–1169. https://doi.org/10.1162/qss_a_00155

  Hsiao, T.-K. (2023). Appropriateness of citing retracted articles in biomedicine: Sentiments expressed in citations without acknowledgement of retraction. *Proceedings of ISSI 2023 – the 19th International Conference of the International Society for Scientometrics and Informetrics, 2*, 181–187. https://doi.org/10.5281/zenodo.8370893

RQ 2a: Citations to retracted articles

- Retraction officially flags problematic science.
- Retraction does not stop the diffusion of the retracted article.
  (Bar-Ilan & Halevi, 2017; Bolland et al., 2021; Candal-Pedreira et al., 2020; Dal-Ré & Ayuso, 2020; Mott et al., 2019; Pfeifer & Snodgrass, 1990; Theis-Mahon & Bakker, 2020; van der Vet & Nijveen, 2016)
  - Previous studies were mostly focused on citation counts or on a particular field.
- The question of why retracted articles continue to be cited after their retractions remains underexplored.
- The current work:
  - A database-wide analysis of citation contexts
  - The purposes of intentionally citing retracted articles
  - Sentiments expressed in post-retraction citation contexts lacking acknowledgement

RQ 2a: Data

- Retracted articles: *"retracted publication" [PT]*
- 7,813 retracted articles
- 48,747 citation contexts (613 excluded)
- Obtain their retraction years by PubMed's *retraction in* links (7,766/7,813; 99.4%)
  - Pre-retraction: 28,439 citation contexts
  - In retraction year: 6,412 citation contexts
  - Post-retraction: 13,252 citation contexts
  - Missing retraction year: 31 citation contexts

RQ 2a: Method

- Identifying intentional post-retraction citations

Priority	Rule	# citation contexts identified	# citation contexts acknowledging the retraction	# false positives
1	At least one of the cue words (retract*, withdr*, and error) appears in the citation context.	243	169	74
2	At least one of the cue words (retract*, withdr*) appears in the acknowledgment window.	309	283	26
3	Retraction notice is cited together with the retracted article in the citing article's full-text.	159	159	0
Total	-	711	611	100

  - *retract** in *neurite retraction*; *withdr** in *withdrawal symptoms*
  - + 111 implicit intentional post-retraction citation contexts
- Annotating citation purposes
  - A classification scheme consisting of 11 categories
- Annotating citation sentiments
  - 25% of post-retraction citation contexts without acknowledgements
  - *strongly positive*, *weakly positive*, *neutral*, and *negative*.

RQ 2a: A model for automatically identifying sentiments

- Customized Bag-of-Words features
  - Sentiments are often hinted at in verbs, adjectives, and adverbs.

    The original report of an alphaproteobacterial Sphingomonas-related GAO (Retracted PMID: 15256569) was later shown to be incorrect and the FISH probes were shown to be binding to members of the Defluviicoccus cluster 1.

  - Part-of-Speech tags
    - Exclude lemmas tagged as proper nouns, numbers, punctuation, and symbols.
    - Only a selected set of nouns were retained.
- Word and sentence embeddings
  - BioWordVec (Zhang et al., 2019) and BioSentVec (Chen et al., 2019)
- Sentiment scores
  - $SentiSentence = \sum_{i=1}^{n} \frac{SentiTerm_i}{D}$
    where $SentiTerm_i$ = 1 for a positive term; $SentiTerm_i$ = -1 for a negative term; D is the distance between the term and the citation marker.
- Pairwise cosine similarity

RQ 2a: Major findings

- Only 5.4% (722/13,252) are the intentional post-retraction citations.

- The findings reported in the retracted articles were still regarded as parts of the development of a particular research topic.

An example of a negative mention (from PMID: 17474991):

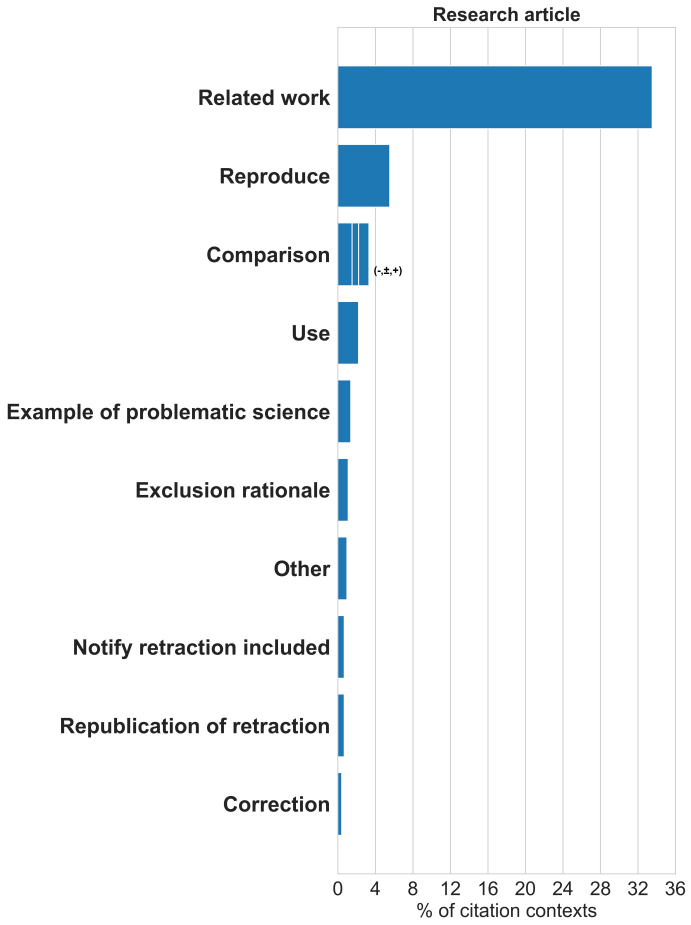
> Another trial of a multivitamin and multimineral supplement in healthy elderly subjects reported beneficial effects after one year in six of seven tests [Retraction PMID: 11527656], though these findings have recently been retracted in the light of concerns about the veracity of the data and possible conflicting commercial interest [Retraction notice PMID: 11527656].

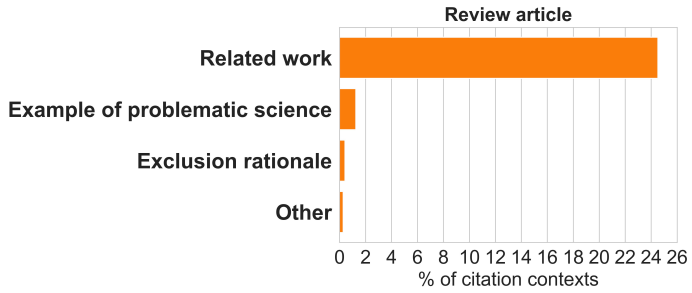
An example of a non-negative mention (from PMID: 26029167):

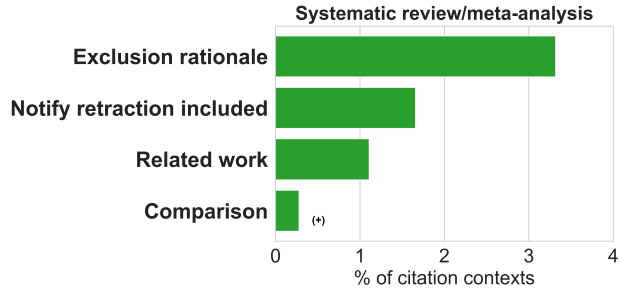
> One of the first biomarkers proposed was serum IGF-I. Despite the retraction of one study suggesting that elevated pre-treatment free IGF-I levels were associated with NSCLC patient response to figitumumab ([Retraction PMID: 21102589]), additional evidence supporting these findings has been published.

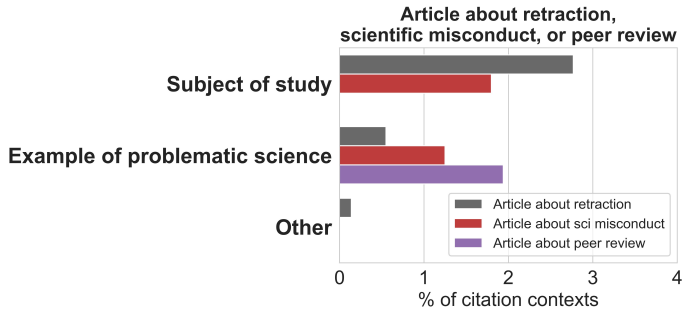
RQ 2a: Major findings

- Sometimes citations to retractions were inevitable.
  - Articles about problematic science
  - Systematic reviews









RQ 2a: Major findings

- Retracted articles were frequently cited as legitimate work without informing the readers about the retraction.

Sentiment	# Citation contexts	(%)
Weakly positive	2,078	(65.84)
Strongly positive	582	(18.44)
Neutral	207	(6.56)
Negative	289	(9.16)
Total	3,156	(100)

- Deep learning models outperformed traditional machine learning models.

Model	Features	CNN		BiLSTM		SVM		Logistic Regression	
		Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
Base model	Word embeddings/BOW*	0.77	0.51	0.78	0.53	0.54	0.43	0.56	0.45
Augmented model 1	Word embeddings/BOW* +Sentence embeddings	0.79	0.60	0.78	0.59	0.65	0.54	0.62	0.52
Augmented model 2	Word embeddings/BOW* +Sentence embeddings +Pairwise cosine similarity +Sentiment scores	0.79	0.60	0.78	0.59	0.65	0.54	0.62	0.53

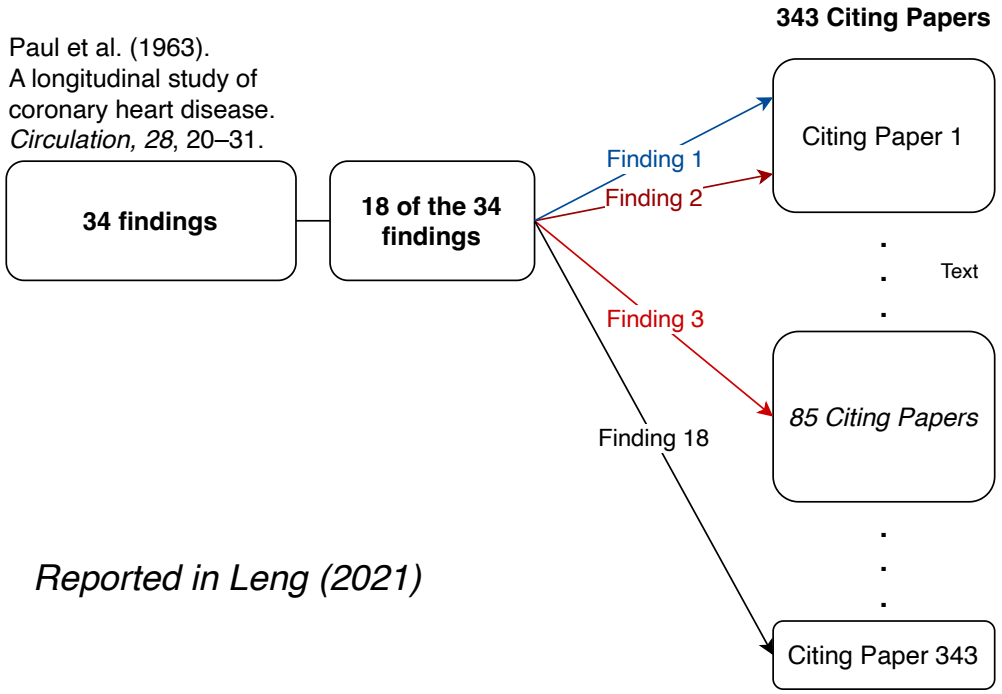
*Word embeddings were used in the CNN and BiLSTM models; BOW was used in the SVM and Logistic Regression models.

RQ 2b - Proposed work: Diversity of citation contexts

"For citations, the cited document is the 'object' and the 'idea' is expressed in the text which cites it."

- Henry G. Small (1978)

- A paper may contain multiple knowledge contents (concepts, methods, or findings)
- Authors may reuse text. (Citron & Ginsparg, 2015; Liu & Chen, 2021)
- Most of the citations mentioned the same knowledge content. (Leng, 2021; Shkurko, 2018; Small, 1978)
- Focused on highly cited papers



Reported in Leng (2021)

RQ 2b - Proposed work: Diversity of citation contexts

“For citations, the cited document is the ’object’ and the ’idea’ is expressed in the text which cites it.”

- Henry G. Small (1978)

RQ 2b - Proposed work: Diversity of citation contexts

“For citations, the cited document is the ’object’ and the ’idea’ is expressed in the text which cites it.”

- Henry G. Small (1978)

When an article receives multiple citations, how diverse are the citation contexts?

RQ 2b: Measuring diversity

- Sentence embeddings
  - A PubMedBERT model released by Deka et al. (2022)
  - BioSimCSE model (Kanakarajan et al., 2022)

- Text diversity metric proposed by Lai et al. (2020)

$$\text{Embedding Diversity} = \sqrt[H]{\prod_{i=1}^{H} \sigma_i}$$

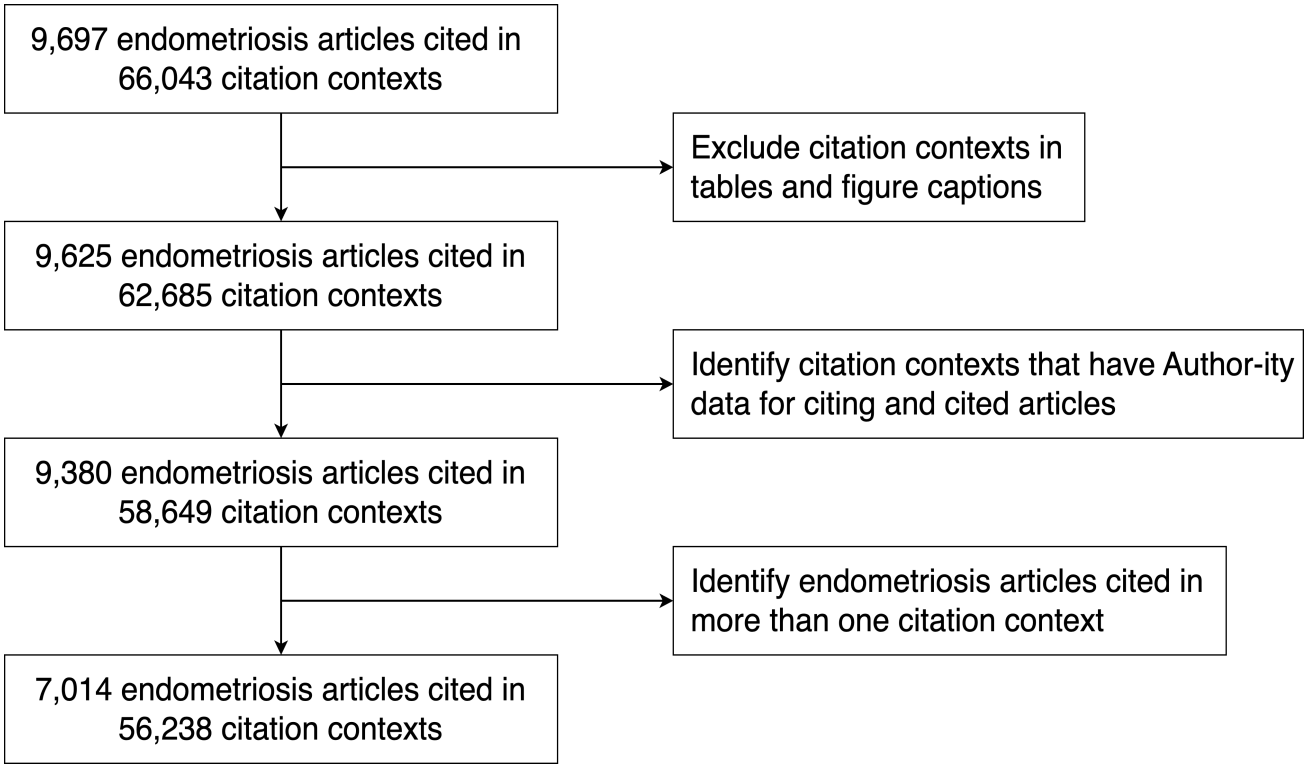
  - σ_i is the standard deviation along a dimension
  - H is the number of dimensions of the sentence embedding.

RQ 2b: Subjects of study

- Endometriosis articles
  - Classically defined as endometrial-like tissues that are present outside the uterus (Amro et al., 2022; Taylor et al., 2021).
  - A research area with ongoing debates and advancements.
    - Used to be considered to arise from retrograde menstruation carrying implants of endometrial tissue outside of the uterus.
    - Recent studies have proposed that genetic-epigenetic incidents can better explain the pathophysiology of endometriosis.
  - The shift in the understanding of endometriosis may be reflected in the citation contexts that an article may be mentioned in different ways.

RQ 2b: Data

- PubMed articles having “endometriosis” in their Medical Subject Headings (MeSH)
- 9,697 of the 20,369 articles were cited in OpCitance



RQ 2b: Preliminary results

Diversity score	PMID	Citation context
0.01	23762683	The incidence of concomitant pelvic endometriosis with scar endometriosis has been reported to be from 14.3% to 26% [cited PMID: 1685456].
	27857904	The incidence of concomitant pelvic endometriosis with scar endometriosis has been reported to be ranging from 14.2% to 26% [cited PMID: 1685456].
0.29	24927773	In fact, immune deficits fulfilling most of the basic criteria for autoimmune disease have been described in endometriosis, including polyclonal B-cell activation, abnormalities in T- and B- cell function, tissue damage, and multi-organ involvement [cited PMID: 20797713 and another article].
	26439741	Given the autoimmune characteristics of endometriosis, endometriosis and nickel allergy may share common features [cited PMID: 20797713].
	29234882	However, it is worth to mention that other class of MHC genes located near HLA-G (HLA- DQ and HLA-DRB1) have already been published in the context of endometriosis [cited PMID: 20797713 and other two articles].
	28611158	Interestingly, genome-wide association studies of human populations have revealed that single-nucleotide polymorphisms in the single human Ccl21 gene are associated with autoimmune diseases, including rheumatoid arthritis and dermatomyositis [cited PMID: 20797713 and other four articles].

RQ 2b - Future work: Embedding diversity metric evaluation

- Relationship between diversity of citation contexts and citation purposes

- The SciCite dataset
  - 11,020 citation contexts with annotations of three categories of citation purposes (Cohan et al., 2019)
    - *Background information*, *Method*, and *Result comparison*
  - 4,914 (44.59%) were from articles with PMIDs.

RQ 2b - Future work: Embedding diversity metric evaluation

- The diversity of biomedical named entities in the citation contexts
  - Biomedical-related nouns and noun phrases
  - Partially reflect the knowledge contents mentioned in the citation contexts
- NCBI's PubTator
- Gini-Simpson index

$$\text{Entity Diversity} = 1\text{-} \sum_{i=1}^{R} \left(\frac{n_i \, (n_i - 1)}{N \, (N - 1)} \right)$$

  - n_i is the count of an entity's appearances in the citation contexts
  - N is the total number of entities' appearances in the citation contexts
  - R is the unique number of entities in the citation contexts

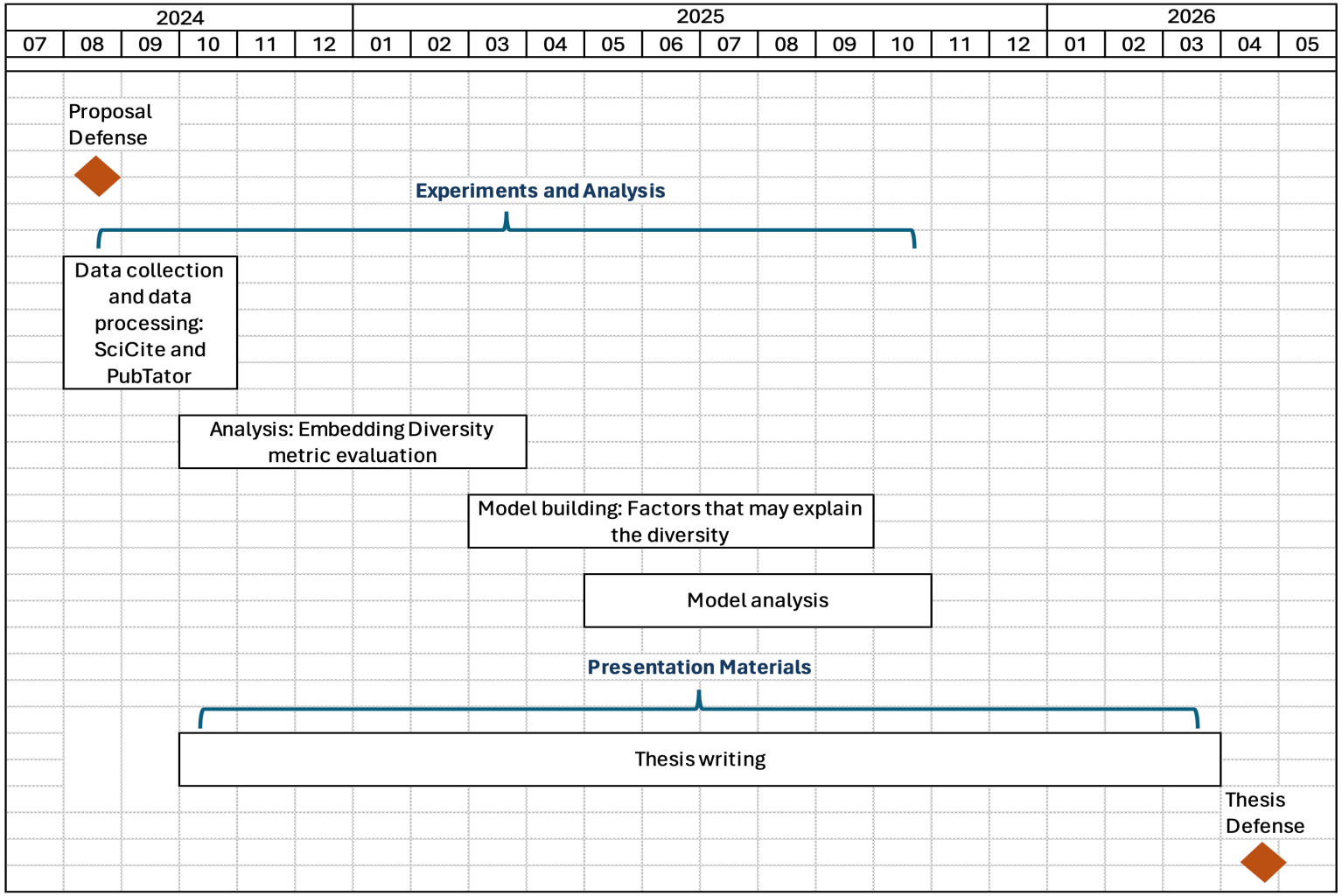
RQ 2b - Future work: Factors that may explain the diversity of citation contexts

- Previous studies focused on the number of distinct knowledge contents in an article that had been cited in the literature.
(Bornmann et al., 2020; Leng, 2021; Shkurko, 2018; Small, 1978)

- Gap: What factors may influence the diversity of citation contexts mentioning an article?

Category	Factor
Author-related	Number of unique citing authors, Diversity of citing authors, Number of unique disciplines of citing authors, Diversity of citing authors' disciplines, Number of citation contexts being self- citations
Article-related	Publication type, Length of the cited article, Years between publication year and the last citation, Number of citing journals
Other	Number of articles mentioned in the citation contexts, Length of the citation contexts

Timeline


2024
2025
2026
07 08 09 10 11 12 01 02 03 04 05 06 07 08 09 10 11 12 01 02 03 04 05
Proposal Defense
Experiments and Analysis
Data collection and data processing: SciCite and PubTator
Analysis: Embedding Diversity metric evaluation
Model building: Factors that may explain the diversity
Model analysis
Presentation Materials
Thesis writing
Thesis Defense

References

Amro, B., Ramirez Aristondo, M. E., Alsuwaidi, S., Almaamari, B., Hakim, Z., Tahlak, M., Wattiez, A., & Koninckx, P. R. (2022). New understanding of diagnosis, treatment and prevention of endometriosis. *International Journal of Environmental Research and Public Health, 19*(11), 6725. https://doi.org/10.3390/ijerph19116725

Bar-Ilan, J., & Halevi, G. (2017). Post retraction citations in context: A case study. *Scientometrics, 113*(1), 547–565. https://doi.org/10.1007/s11192-017-2242-0

Bolland, M. J., Grey, A., & Avenell, A. (2021). Citation of retracted publications: A challenging problem. *Accountability in Research, 29*(1), 18–25. https://doi.org/10.1080/08989621.2021.1886933

Bornmann, L., Wray, K. B., & Haunschild, R. (2020). Citation concept analysis (CCA): A new form of citation analysis revealing the usefulness of concepts for other researchers illustrated by exemplary case studies including classic books by Thomas S. Kuhn and Karl R. Popper. *Scientometrics, 122*(2), 1051–1074. https://doi.org/10.1007/s11192-019-03326-2

Candal-Pedreira, C., Ruano-Ravina, A., Fernández, E., Ramos, J., Campos-Varela, I., & Pérez-Ríos, M. (2020). Does retraction after misconduct have an impact on citations? A pre–post study. *BMJ Global Health, 5*(11), Article 11. https://doi.org/10.1136/bmjgh-2020-003719

Chen, Q., Peng, Y., & Lu, Z. (2019). BioSentVec: Creating sentence embeddings for biomedical texts. *2019 IEEE International Conference on Healthcare Informatics*, 1–5. https://doi.org/10.1109/ICHI.2019.8904728

Citron, D. T., & Ginsparg, P. (2015). Patterns of text reuse in a scientific corpus. *Proceedings of the National Academy of Sciences, 112*(1), 25–30. https://doi.org/10.1073/pnas.1415135111

Cohan, A., Ammar, W., van Zuylen, M., & Cady, F. (2019). Structural scaffolds for citation intent classification in scientific publications. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 3586–3596. https://doi.org/10.18653/v1/N19-1361

Dal-Ré, R., & Ayuso, C. (2020). For how long and with what relevance do genetics articles retracted due to research misconduct remain active in the scientific literature. *Accountability in Research, 28*(5), 280–296. https://doi.org/10.1080/08989621.2020.1835479

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics, 6*, 391–406. https://doi.org/10.1162/tacl_a_00028

Khadka, A., & Knoth, P. (2018). Using citation-context to reduce topic drifting on pure citation-based recommendation. *Proceedings of the 12th ACM Conference on Recommender Systems*, 362–366. https://doi.org/10.1145/3240323.3240379

Leng, R. I. (2021). Diversity in citations to a single study: A citation context network analysis of how evidence from a prospective cohort study was cited. *Quantitative Science Studies, 2*(4), 1216–1245. https://doi.org/10.1162/qss_a_00154

Liu, Y., & Chen, M. (2021). Applying text similarity algorithm to analyze the triangular citation behavior of scientists. *Applied Soft Computing, 107*, 107362. https://doi.org/10.1016/j.asoc.2021.107362

Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. (2020). S2ORC: The Semantic Scholar Open Research Corpus. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4969–4983. https://doi.org/10.18653/v1/2020.acl-main.447

Mott, A., Fairhurst, C., & Torgerson, D. (2019). Assessing the impact of retraction on the citation of randomized controlled trial reports: An interrupted time-series analysis. *Journal of Health Services Research & Policy, 24*(1), Article 1. https://doi.org/10.1177/1355819618797965

Pfeifer, M. P., & Snodgrass, G. L. (1990). The continued use of retracted, invalid scientific literature. *JAMA, 263*(10), 1420-1423. https://doi.org/10.1001/jama.1990.03440100140020

Shkurko, A. V. (2018). How many knowledge claims are there in a scientific text? A study of three neuroscientific articles' content as reflected in the citing publications. *Sociology of Science and Technology, 9*(2), 71–85. https://doi.org/10.24411/2079–0910–2018–10005

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science, 8*(3), 327–340. https://doi.org/10.1177/030631277800800305

Smith, L. C. (1981). Citation analysis. *Library Trends, 30*(1), 83–106.

Taylor, H. S., Kotlyar, A. M., & Flores, V. A. (2021). Endometriosis is a chronic systemic disease: Clinical challenges and novel innovations. *The Lancet, 397*(10276), 839–852. https://doi.org/10.1016/S0140-6736(21)00389-5

Theis-Mahon, N. R., & Bakker, C. J. (2020). The continued citation of retracted publications in dentistry. *Journal of the Medical Library Association, 108*(3), 389–397. https://doi.org/10.5195/jmla.2020.824

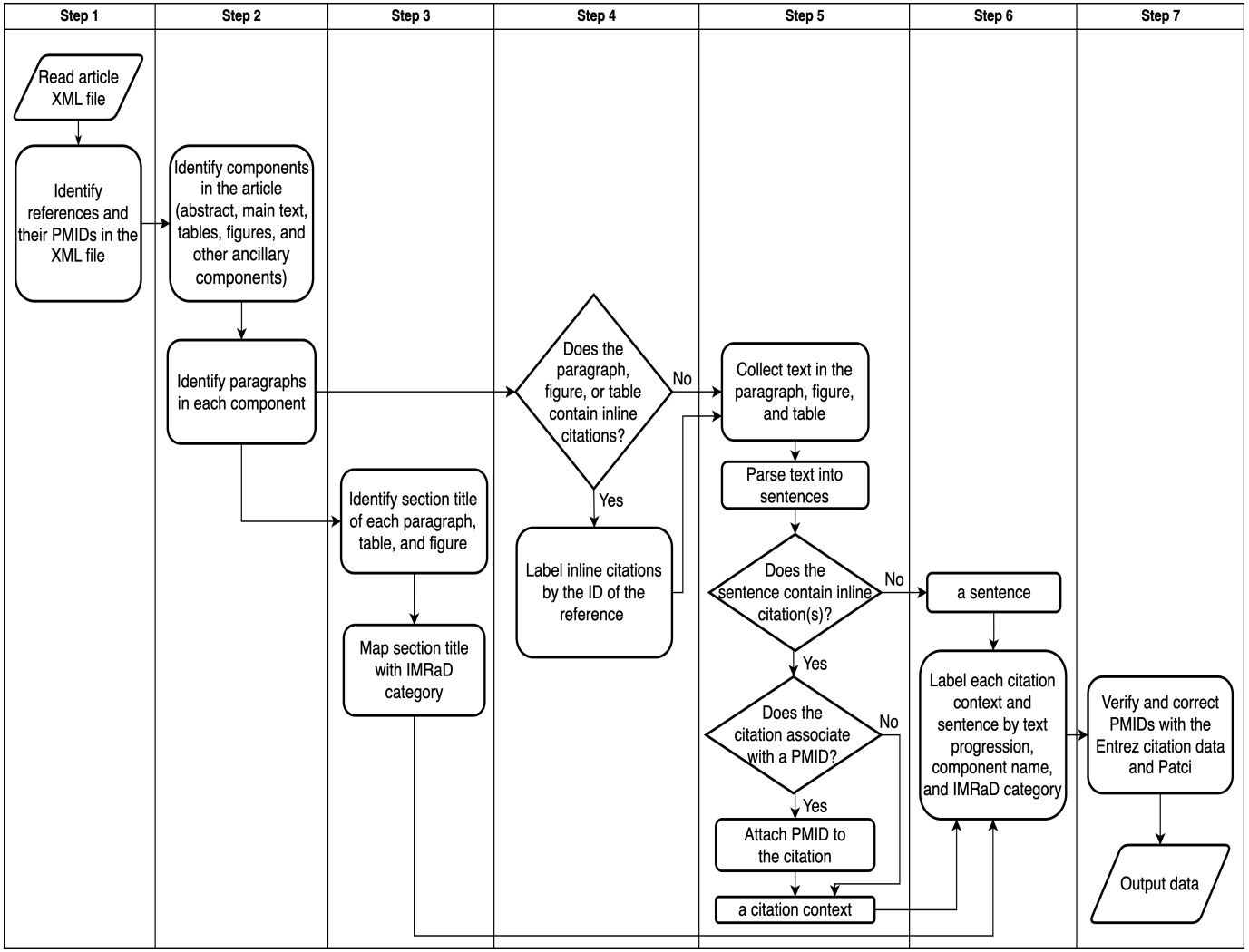
van der Vet, P. E., & Nijveen, H. (2016). Propagation of errors in citation networks: A study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal Nature. *Research Integrity and Peer Review, 1*, 3. https://doi.org/10.1186/s41073-016-0008-5

Wu, J., Hiltabrand, R., Soós, D., & Giles, C. L. (2022). Scholarly big data quality assessment: A case study of document linking and conflation with S2ORC. *Proceedings of the 22nd ACM Symposium on Document Engineering*, 1–4. https://doi.org/10.1145/3558100.3563850

Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data, 6*(1), 52. https://doi.org/10.1038/s41597-019-0055-0

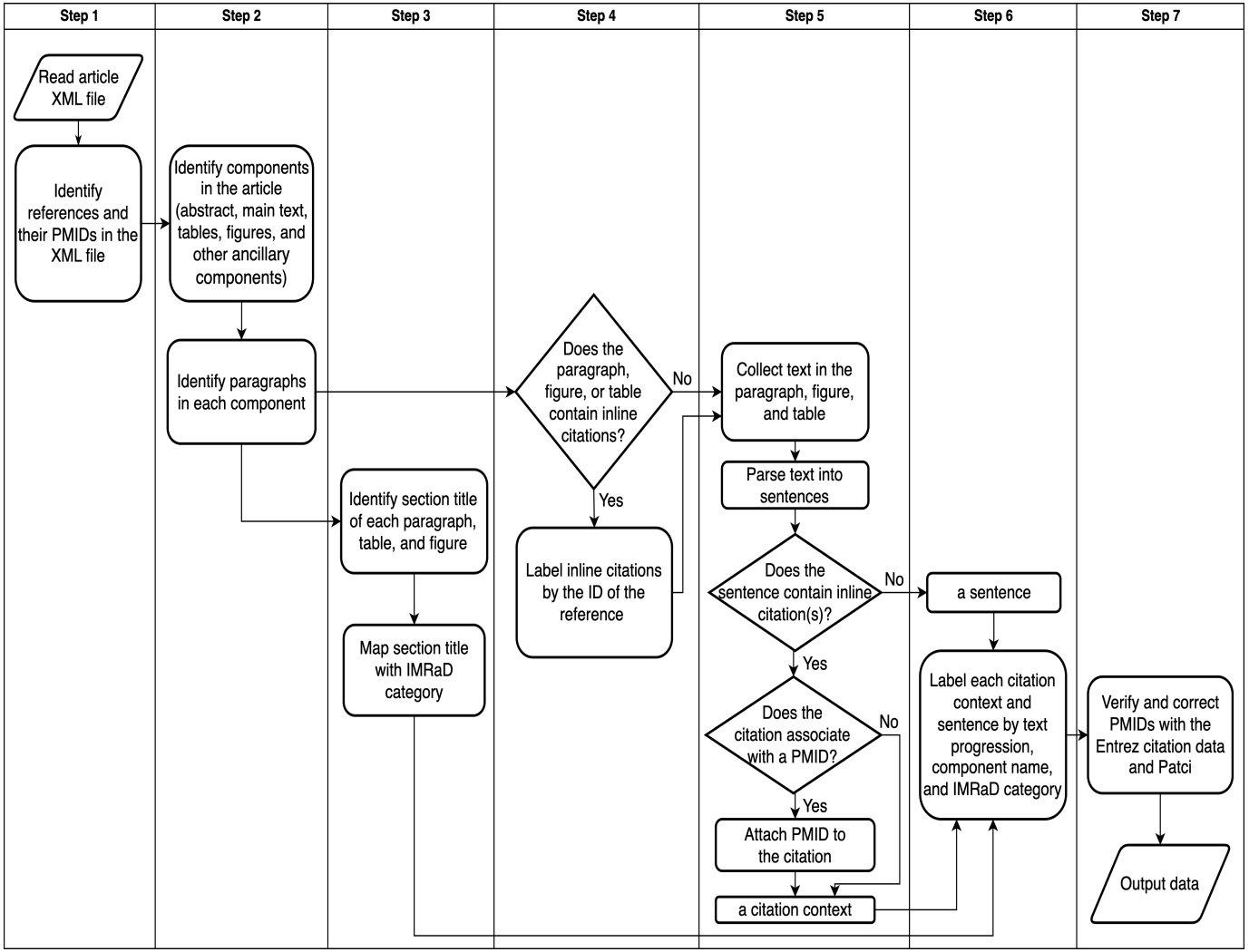
OpCitance: The pipeline

- References and their PMIDs
  (`ref` tags; `pub-id` tags)



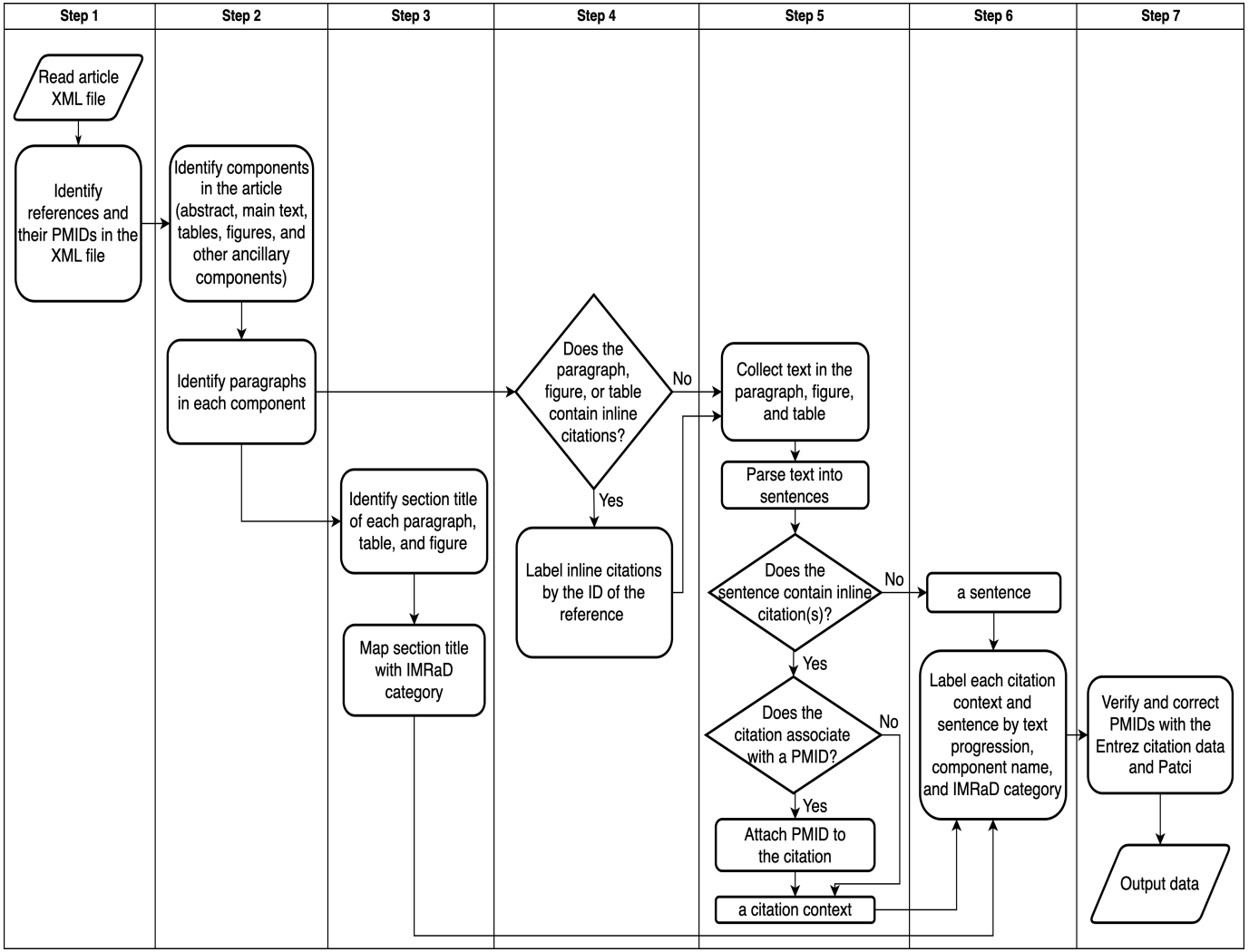
OpCitance: The pipeline

- References and their PMIDs (`ref` tags; `pub-id` tags)
- Components in an article



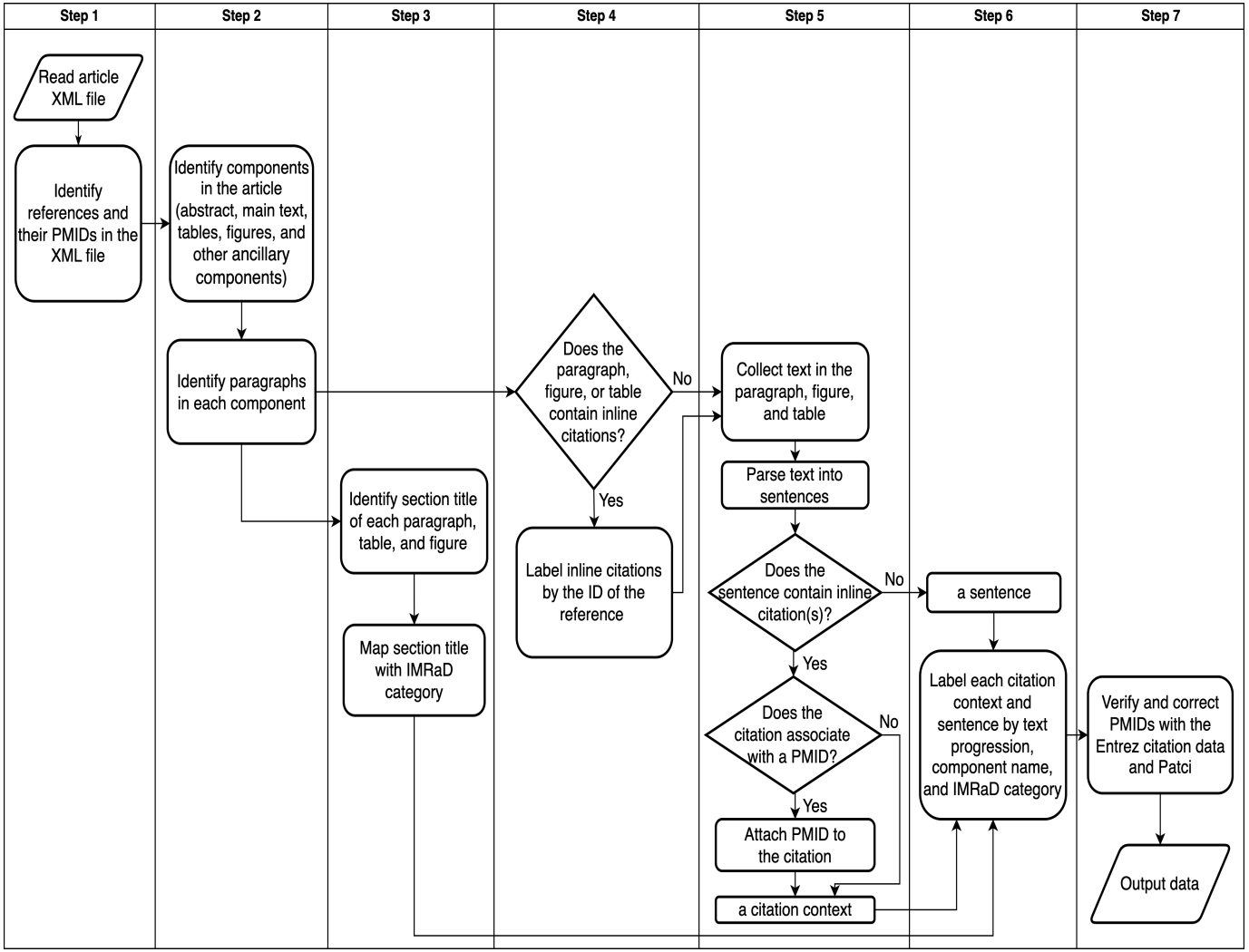
OpCitance: The pipeline

- References and their PMIDs
  (`ref` tags; `pub-id` tags)
- Components in an article
- IMRaD
  (section titles and section types)



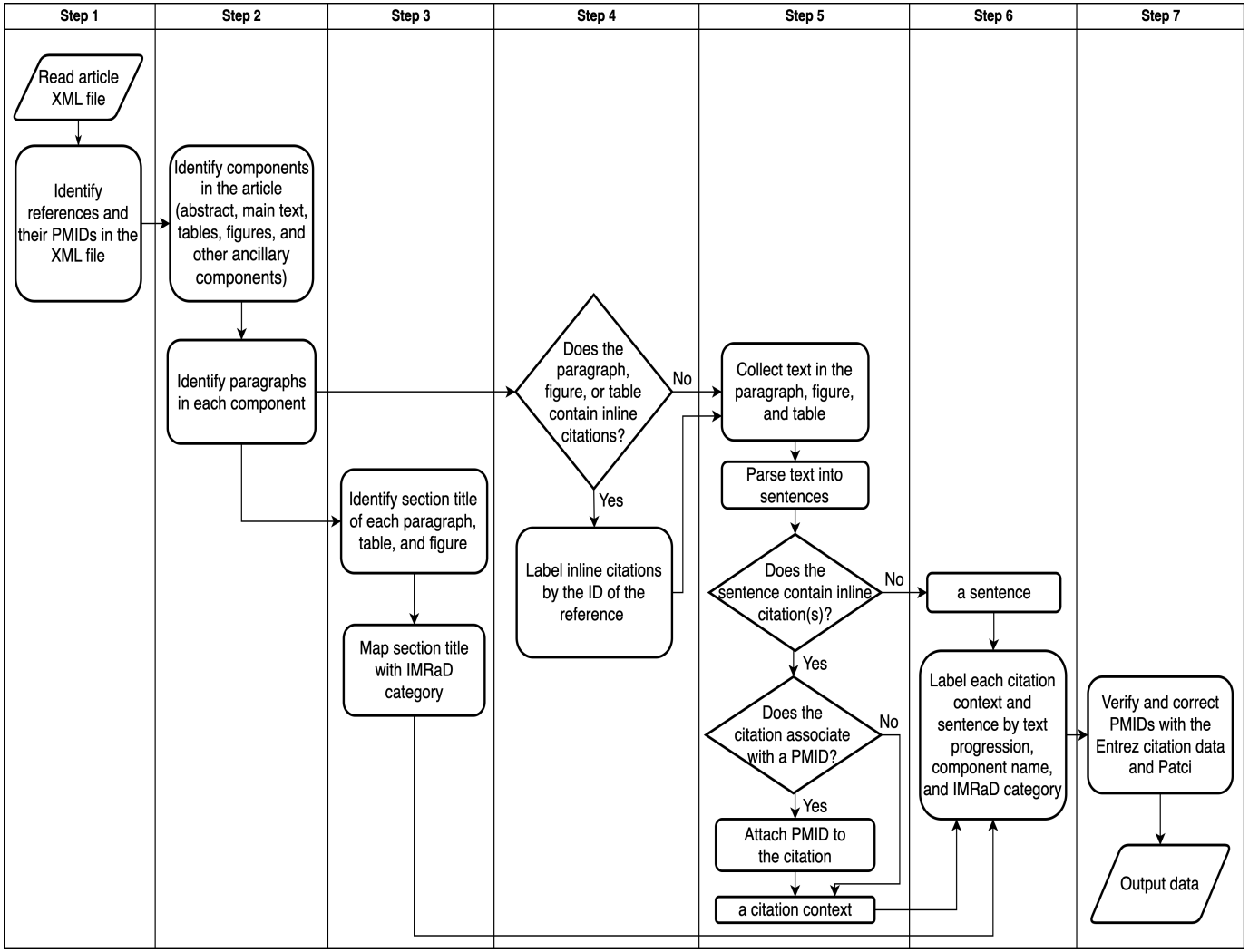
OpCitance: The pipeline

- References and their PMIDs
  (`ref` tags; `pub-id` tags)
- Components in an article
- IMRaD
  (section titles and section types)
- Inline citations
  (`xref` tags)



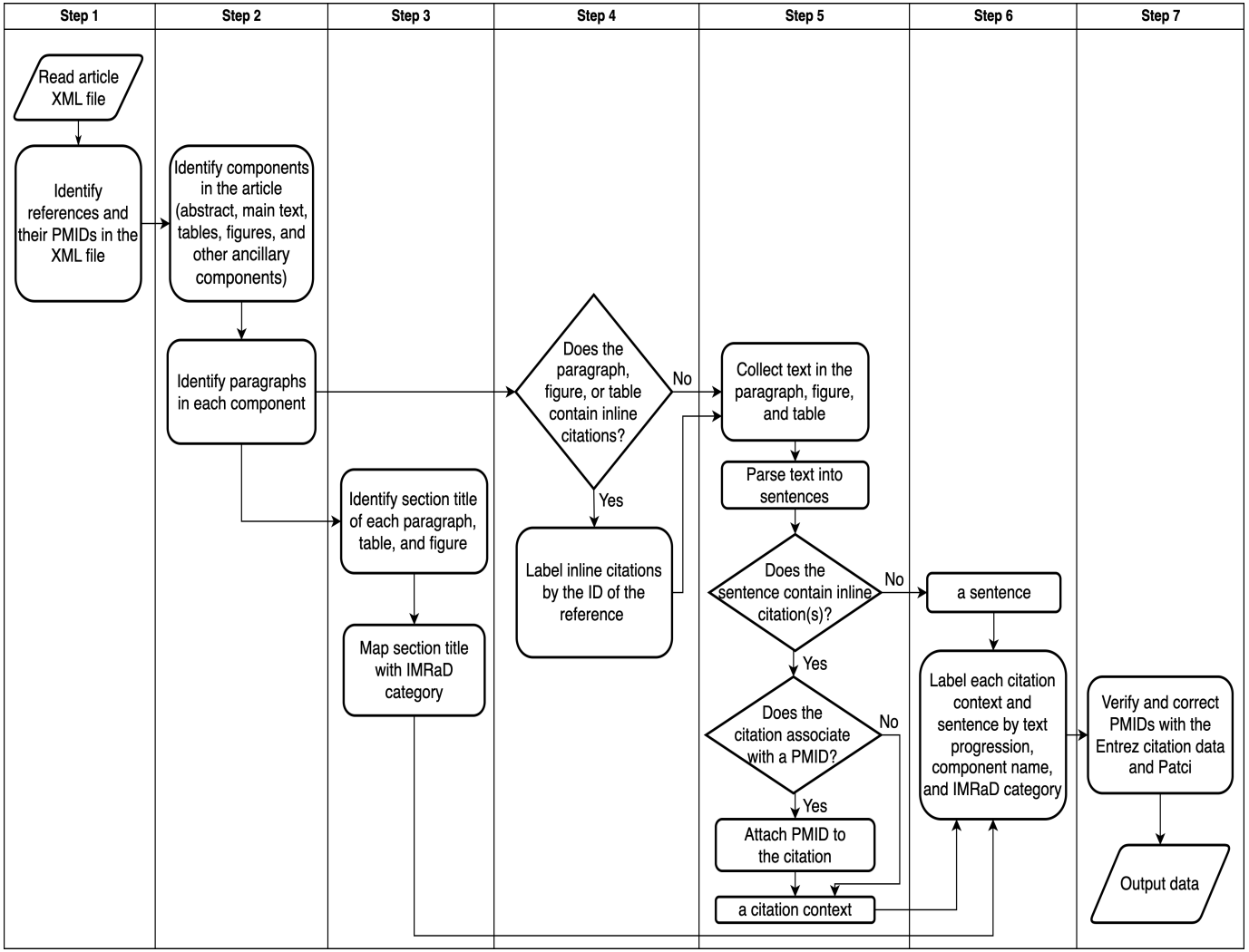
OpCitance: The pipeline

- References and their PMIDs
  (`ref` tags; `pub-id` tags)
- Components in an article
- IMRaD
  (section titles and section types)
- Inline citations
  (`xref` tags)
- Citation contexts



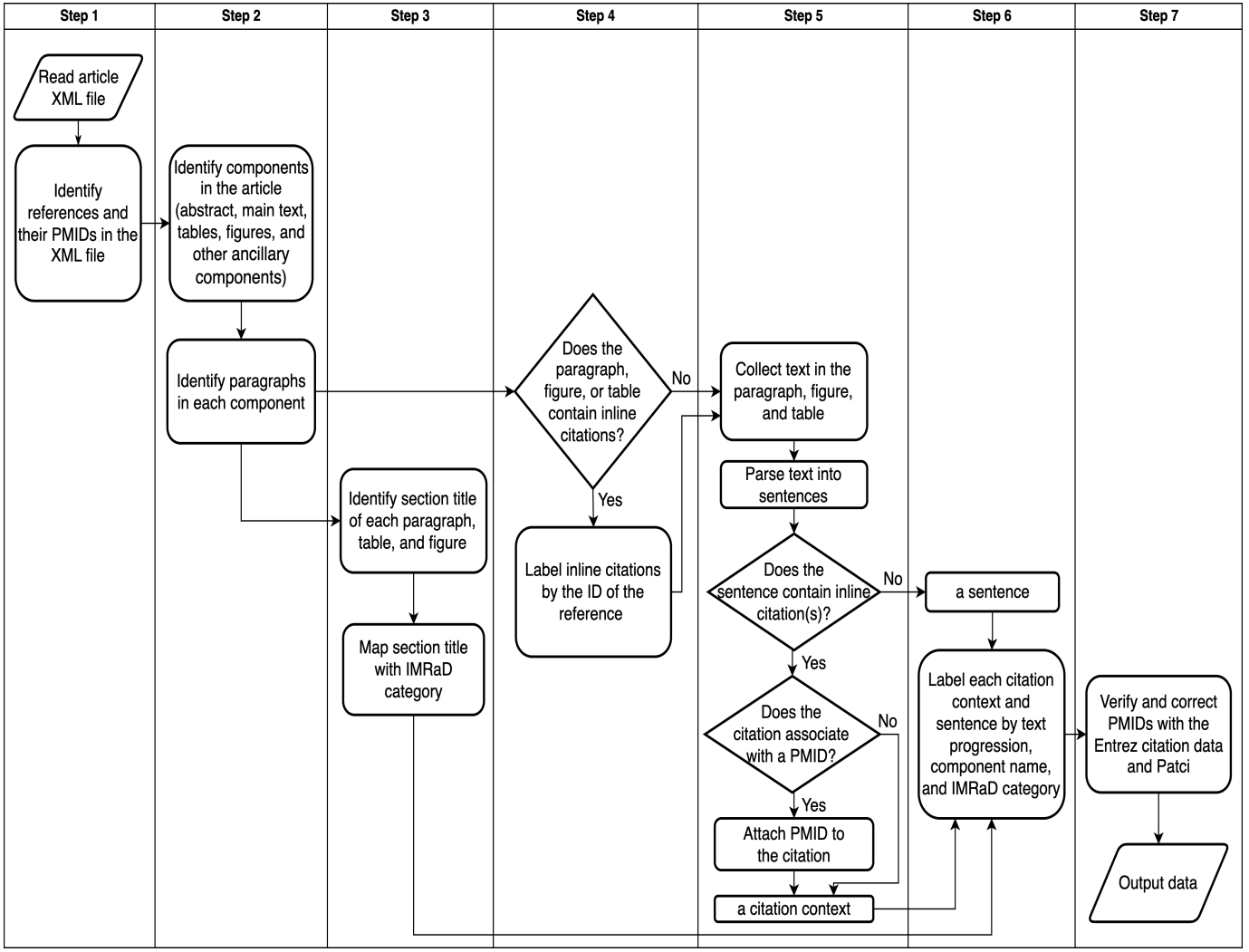
OpCitance: The pipeline

- References and their PMIDs
  (`ref` tags; `pub-id` tags)
- Components in an article
- IMRaD
  (section titles and section types)
- Inline citations
  (`xref` tags)
- Citation contexts
- Labelling sentences



OpCitance: The pipeline

- References and their PMIDs
  (`ref` tags; `pub-id` tags)
- Components in an article
- IMRaD
  (section titles and section types)
- Inline citations
  (`xref` tags)
- Citation contexts
- Labelling sentences
- Verify XML-tagged PMIDs



OpCitance: Identify IMRaD categories

- Section titles
  - `sec` tags
  - `title` tags or `label` tags
- Section types
  - `sec-type` attribute

Label	IMRaD category	Cue words and phrases
I	Introduction/Background	intro*, overview, background, history, related work, related stud*, previous work, previous stud*, review
M	Method	method, material, experimental procedure, protocol, data
R	Result	result, finding
D	Conclusion/Discussion	conclud*, conclusion, summary, discuss*, future
NoIMRaD	-	The string does not contain the above terms.

OpCitance: Labelling sentences

- Label by their components (abstract, body, etc.)
- IMRaD only applied to body text

PMCID	Location	IMRaD	Sentence
5675298	abstract	NoIMRaD	Information about benign prostatic hyperplasia (BPH) has become increasingly accessible on the Internet. Though the ability to find such material is encouraging, its readability and impact on informing patient decision making are not known.
5675298	body	I	Benign prostatic hyperplasia (BPH) is a prevalent condition affecting the quality of life of millions of adult men (|bibr38-1557988316680935|).
5675298	appendix	NoIMRaD	Websites at or Below the Eighth-Grade SMOG Reading Level.

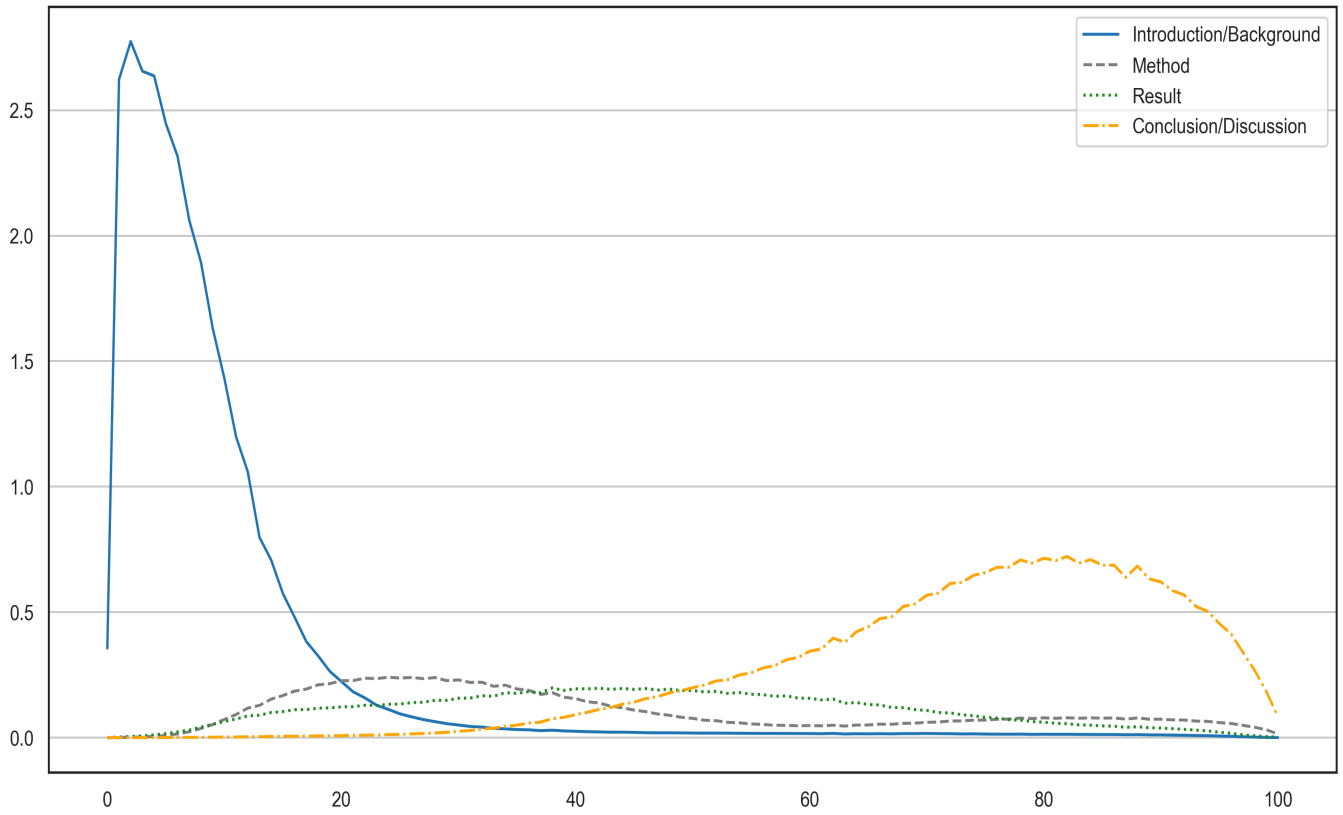
OpCitance: Quality evaluation (document linking)

- Entrez citation data
  - 70.96% (1,290,693 out of 1,818,893) of the articles had at least one discrepancy between the PMIDs of citations.
- Patci
  - 98.25% (135,340,795/137,748,787) of the inline citations were verified.

Indicator value	Definition	# Inline citations (%)
SAME	The Patci-identified ID is the same as the XML- tagged PMID.	101,885,318 (75.28)
NONE	The citation has neither a Patci-identified ID nor an XML-tagged PMID.	21,452,327 (15.85)
INSERT	The citation has a Patci-identified ID but lacks an XML-tagged PMID.	11,595,741 (8.57)
SWAP	The Patci-identified PMID is different from the XML-tagged PMID.	317,513 (0.23)
DELETE	The citation has an XML-tagged PMID, but Patci does not identify any ID for it.	89,896 (0.07)

OpCitance: Quality evaluation (IMRaD)

- Inline citations in IMRaD sections by text progression



OpCitance: Quality evaluation (IMRaD)

- Inline citations in IMRaD sections by text progression
- Precision, recall, and F1

OpCitance: Quality evaluation (IMRaD)

- Inline citations in IMRaD sections by text progression
- Precision, recall, and F1

Evaluation Level	IMRaD	Precision	Recall	F1
Section	I	0.957	0.854	0.903
	M	1.000	0.820	0.901
	R	1.000	0.915	0.956
	D	1.000	0.993	0.996
	NoIMRaD	0.638	0.949	0.763
	Macro average	0.919	0.905	0.904

RQ 2a: Citation purposes and examples

Purpose	Description	Example
Comparison	Authors of the citing paper compared "their" results or methods with the retracted paper. According to the tone, this category is further divided into negative (-), positive (+), and neutral (±). The negative tone refers to the cases that inconsistency, contradiction, or discrepancy is reported in the comparison. The positive tone refers to the cases that consistency is reported in the comparison. The neutral tone refers to the cases that the consistency between the compared results was unclear.	Overall, our results provide no evidence for a beneficial effect of multivitamin and multimineral supplementation on cognitive function in the majority of men and women 65 years and over living in the community. This result is consistent with all previous studies in non-selected elderly populations [2-10] apart from the retracted Canadian study [Retraction PMID:11527656].
Correction	The retracted paper was cited to make a correction.	In our previous publication [1], Figure 4 involved the analysis of chemotherapy-response signatures (as carried out independently by author AP and described in a 2006 Nature Medicine article [Retraction PMID: 17057710]). It has recently been determined that the chemotherapy-response signatures in [Retraction PMID: 17057710] are not reproducible, causing retraction of that article. As such, the results presented in Figure 4 of our original paper [1] are no longer valid.
Example of problematic science	The retracted paper was cited to provide an example of problematic science. This purpose satisfies one of the following conditions: (1) The retracted paper was cited to provide an example of problematic research (e.g., irreproducible research, unreliable research, research involving scientific misconduct, a flawed study, etc.). (2) The retracted paper was cited to provide an example where peer review failed, and problematic science was published. (3) The retracted paper was cited to provide an example showing a problem in scientific research or scholarly communication. (4) The retracted paper was cited to provide an example of the societal impact of problematic research.	There are also cases of data manipulation/fraud in NIH-supported, peer-reviewed research—more than 30 cases documented in the past 3 years according to the U.S. Public Health Service's Office of Research Integrity (2009). One example is the study of Arnold et al. [Retraction PMID: 8633243] in which the authors reported huge synergistic effects of endocrine disruptors in the yeast estrogen assay in vitro. McLachlan [Retraction notice PMID:9254413] rescinded that paper because neither his laboratory nor others could replicate the findings. It was later determined that there was scientific misconduct and the original data were fabricated (NIH 2001).

RQ 2a: Citation purposes and examples (continued)

Purpose	Description	Example
Exclusion rationale	The retracted paper was cited to explain why it is excluded from use/consideration. Especially found in the context of research synthesis (e.g. review articles and meta- analyses which provide a formal exclusion rationale for papers that are not included.) This purpose can also be found in the literature review section of a research article.	The meta-analysis presented here provides a valid and up to date summary of the relevant literature, including a recently published randomised controlled trial of 339 patients with a confirmed diagnosis of uncomplicated appendicitis. It excludes the study that has been retracted subsequent to publication [Retraction PMID: 19277796], as well as another for which it was not clear if patients were randomised.
Notify retraction included	Notify readers that one or more retracted papers were included in a different, previous published review article, guideline, or paper.	Systematic reviews and meta-analyses of the trials, including a Cochrane review comparing antibiotic treatment and appendicectomy, published in recent years summarised the evidence as either in favour of antibiotic treatment or inconclusive. This could possibly result from inclusion of trials with poor methods or retracted since publication [Retraction PMID: 19277796], or from simplifying the evidence as a summary of both randomised and non-randomised studies.
Related work	The retracted paper was cited to show what has been done or found in the past or was cited for one of the following reasons: (1) The Retracted paper was once a landmark in the field; (2) the retracted paper was the origin/pioneer of something (e.g., "X first identified/describe Y", "X was identified as a novel ....", "X was initially proposed...", or "X was originally..."); (3) the retracted paper led to an important event in the field, such as Wakefield's paper's influence on the autism-vaccine link and the anti-vaccine movement.	The COOPERATE study [Retraction PMID: 12531578] even showed that dual therapy with trandolapril and losartan reduced the risk of the primary endpoint (time to doubling of serum creatinine level or end stage renal disease) by 60% better than monotherapy, thereby becoming one of the most widely quoted studies by the Lancet. After such seemingly robust evidence many physicians accepted that reduction of albuminuria or proteinuria was synonymous with nephroprotection.
Republication of retraction	In the republication of the retracted paper, the authors cited the retracted paper to announce the republication.	This article is a revised version of a paper of the same title [Retraction PMID: 22241970] that was previously published in PLOS Computational Biology and was subsequently retracted when a computational error was discovered.

RQ 2a: Citation purposes and examples (continued)

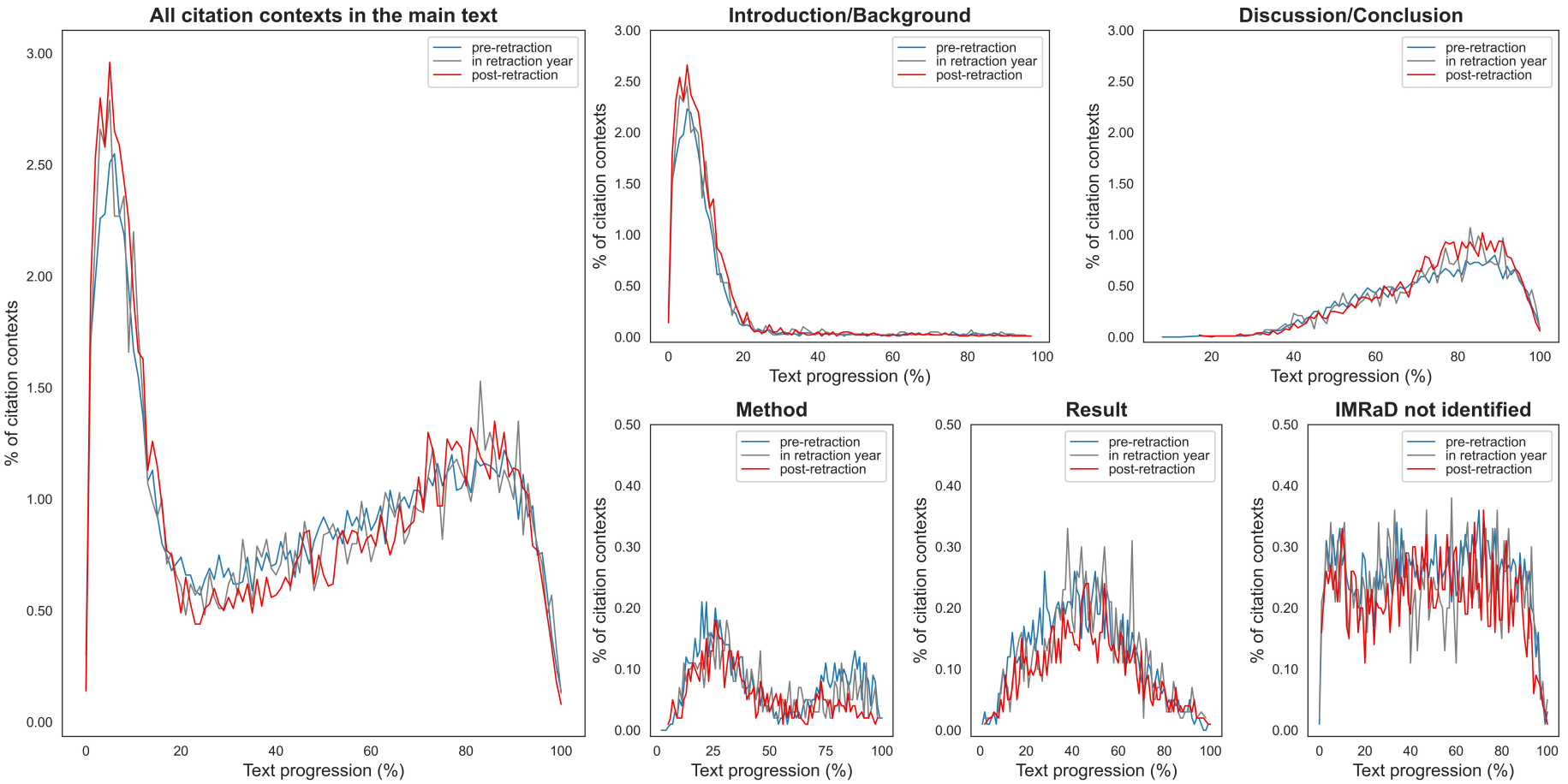
Purpose	Description	Example
Reproduce	A citation to the retracted paper was made because the citing paper tried to reproduce/repeat the finding or experiment mentioned in the retracted paper.	In a further effort to obtain an intelligence priming effect, we attempted to replicate Gordijn and Stapel's study. Their article [Retraction PMID: 17393877] was retracted after the completion of Experiment 7 and hence their data can be given no evidential weight. Nevertheless, the hypothesis they put forward is a reasonable one and thus we report Experiment 7 in relation to that hypothesis.
Subject of study	Cited retraction is the object of study of a case study about retraction, or is the data used in a study about retraction, scientific misconduct, or peer review. Note that in these studies, retracted papers can be cited in the results.	Our objective was to test whether such simple tools applied to a manuscript known to be fraudulent [Retraction PMID: 19923501] would have helped to detect some warning signals of poor quality.
Use	Citing paper uses something from the cited retracted paper. This type of citation is often found in Methods sections.	To verify whether the above-mentioned points have the potential of playing substantial roles in explaining this story, we planned a proof-of-concept survey of DNA sequence databases, and namely: (i) A bioinformatic analysis on the primer sequences described by Lombardi et al ([Retraction PMID: 19815723]), compared against human and murine (house mouse Mus musculus) genomes.
Other	Those do not belong to the above categories.	--

RQ 2a: Citation sentiments

Sentiment	Definition	Example
Strongly positive	Citing work uses something (e.g., definitions, concepts, materials, equipment, and techniques) in the retracted work. Citing work confirms, is supported by, depends on, or explicitly agrees with the retracted work.	Similar to our results, SLA mRNA expression levels were previously reported to be downregulated in human B-cells and were strongly expressed in naïve, pre-germanal center, and germinal-center B-cells based on gene expression analysis. [Retraction PMID: 12438421].
Weakly positive	Retracted work is cited as related and legitimate, without raising concerns.	Altered NPM1 expression was observed in many types of tumors, and mutated NPM1 is frequently detected in human hematopoietic malignancies, especially in acute myeloid leukemia (AML) [Retraction PMID: 18401421].
Negative	Citing work disputes, corrects, questions, or disagrees with the retracted work. Citing work expresses concerns or casts doubts on the retracted work. Citing work indicates that the retracted work has shortcomings or uncertainty. Citing work mentions that the findings reported in the retracted work are controversial or lacking confirmation.	The original report of an alphaproteobacterial Sphingomonas-related GAO [Retraction PMID: 15256569] was later shown to be incorrect and the FISH probes were shown to be binding to members of the Defluviicoccus cluster 1.
Neutral	Retracted work is cited without additional comment. No judgment of validity is shown in the citation context.	The contribution of inflammatory cytokines to tumor development has been investigated by other studies [Retraction PMID: 25544369, Non-retraction PMIDs: 18954521, 18036640]

RQ 2a: Locations of citations to retracted articles

- No substantial difference between how the retracted articles were cited before and after the retractions.



RQ 2a: Locations of citations to retracted articles

- Single mention

IMRaD	# Pre-retraction pairs	(%)	# Post-retraction pairs (retraction not acknowledged)	(%)	# Post-retraction pairs (retraction acknowledged)	(%)
Introduction/background	3,272	(27.52)	2,359	(33.04)	47	(16.43)
Methods	690	(5.8)	397	(5.56)	13	(4.55)
Results	971	(8.17)	490	(6.86)	29	(10.14)
Discussion/conclusion	3,656	(30.75)	2,349	(32.90)	46	(16.08)
IMRaD not identified	3,300	(27.76)	1,545	(22.64)	151	(52.80)
Total	11,889	(100)	7,140	(100)	286	(100)

RQ 2a: Locations of citations to retracted articles

- Multiple mentions

# Different IMRaD sections	# Pre-retraction pairs	(%)	# Post-retraction pairs (retraction not acknowledged)	(%)	# Post-retraction pairs (retraction acknowledged)	(%)
1	2,225	(44.45)	860	(46.66)	68	(50.00)
2	2,188	(43.71)	890	(48.29)	57	(41.91)
3	483	(9.65)	82	(4.45)	10	(7.35)
4	109	(2.18)	10	(0.54)	1	(0.74)
5	1	(0.02)	1	(0.05)	0	(0)
Total	5,006	(100)	1,843	(100)	136	(100)

RQ 2a: Numbers and percentages of citation contexts by each citation purpose

Purpose	# Citation contexts	(%)
Related work	453	(62.74)
Example of problematic science	62	(8.59)
Reproduce	40	(5.54)
Exclusion rationale	35	(4.85)
Subject of study	33	(4.57)
Comparison	26	(3.60)
Notify retraction included	24	(3.32)
Use	20	(2.77)
Other	14	(1.94)
Correction	10	(1.39)
Republication of retraction	5	(0.69)
Total	722	(100)

RQ 2a: A model for automatically identifying sentiments

- Nouns included in the BOW features:
  - accordance, addition, agreement, analysis, author, concern, contrast, controversy, correlation, example, experiment, evidence, failure, finding, hypothesis, improvement, instance, limitation, literature, method, model, other, report, research, result, study, suspicion, work.

RQ 2a: A model for automatically identifying sentiments

- Best model performance per category

Model	Features	Overall		Per category			
		Accuracy	Macro F1	Sentiment	Precision	Recall	F1
Augmented CNN model 1	Word embeddings +Sentence embeddings	0.79	0.60	Negative	0.61	0.53	0.56
				Neutral	0.40	0.21	0.27
				Weakly positive	0.84	0.92	0.87
				Strongly positive	0.77	0.66	0.70
Augmented CNN model 2	Word embeddings +Sentence embeddings +Pairwise cosine similarity +Sentiment scores	0.79	0.60	Negative	0.62	0.53	0.57
				Neutral	0.38	0.22	0.26
				Weakly positive	0.84	0.91	0.88
				Strongly positive	0.74	0.67	0.70

RQ 2b - Future work: Embedding diversity metric evaluation

Entities (E_i)	Counts of Entities' Appearances in Citation Contexts (n_i)	Total (N)	Calculation	Entity Diversity
$\{E_1, E_2, E_3\}$	{3, 3, 3}	9	$1 - \left(\frac{(3\times2)+(3\times2)+(3\times2)}{9\times8}\right)$	0.75
	{5, 2, 2}	9	$1 - \left(\frac{(5\times4)+(2\times1)+(2\times1)}{9\times8}\right)$	0.67
$\{E_1, E_2, E_3, E_4\}$	{3, 3, 3, 3}	12	$1 - \left(\frac{(3\times2)+(3\times2)+(3\times2)+(3\times2)}{12\times11}\right)$	0.82
	{5, 3, 2, 2}	12	$1 - \left(\frac{(5\times4)+(3\times2)+(2\times1)+(2\times1)}{12\times11}\right)$	0.77