

Towards Claim-Centered Biomedical Knowledge Representation: A Pilot Annotation Corpus

Rohitha Ravinder^{1,2,*,†}, Leyla Jael Castro² and Dietrich Rebholz-Schuhmann^{1,2,†}

¹University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany

²ZB MED - Information Center for Life Sciences, Gleueler Str. 60, 50931 Cologne, Germany

Abstract

Biomedical relation extraction pipelines produce triples that discard the epistemic context and primary findings of scientific literature. We present a pilot annotation corpus that enriches the BioRED corpus with claim centric annotations. These annotations capture the primary scientific claim of each abstract alongside its type, polarity and assertion strength while assigning relation triples a specific semantic role relative to that main claim. We define the annotation dimensions and present a baseline corpus of 150 annotated abstracts to support future claim-centered knowledge representation.

Keywords

claim annotation, biomedical knowledge graphs, knowledge representation, scientific claims

1. Introduction and Motivation

Biomedical knowledge graphs are increasingly used for literature exploration, drug repurposing and clinical decision support [1, 2]. Existing biomedical knowledge graph construction pipelines rely largely on relation extraction (RE) systems that transform scientific literature into structured triples. While such representations capture relational structure, they flatten the epistemic, contextual and provenance properties of scientific findings into context-independent assertions [3].

Biomedical literature contains established facts and novel scientific claims. The primary scientific contribution (main claim) of an abstract is typically a unified, high-level statement supported by multiple underlying observations. Current RE paradigms fail to identify or encode this main claim as a distinct semantic unit. Instead, they extract a pool of binary relation triples, treat background claims, methodological context and primary discoveries identically i.e., they do not distinguish novel claims from background ones. Furthermore, the downstream interpretation of these triples depends on their explicit assertion context, including the author's degree of epistemic commitment, structural polarity and provenance. By extracting relations in isolation, the main claim is fragmented and its context is lost.

We argue that biomedical knowledge representation requires an intermediate semantic layer between relation extraction and knowledge graph construction, representing the paper's primary scientific contribution as a structured semantic object with its own epistemic properties and provenance, rather than as isolated triples.

This paper presents ongoing work towards an annotation grounded in the BioRED corpus [4]. We develop an annotation schema that explicitly models the abstract's primary scientific contribution and its contextual boundaries. Specifically, our framework introduces a claim-centric annotation overlay that: (1) isolates the exact textual span and a normalized free-text formulation of the abstract's main claim, (2) characterizes its high-level epistemic properties (such as type, polarity and assertion strength), and (3) contextualizes the existing BioRED relation triples by assigning each a explicit semantic role relative to that central claim. We present a pilot annotation effort over a sampled target corpus of 150

Posters, Demos, Blue Sky, and Tutorials at SEMANTiCS 2026, Sep 2026, Ghent, Belgium

*Corresponding author.

†These authors contributed equally.

✉ ravinder@zbmed.de (R. Ravinder)

ORCID 0009-0004-4484-6283 (R. Ravinder); 0000-0001-7116-9338 (L. J. Castro); 0000-0002-1018-0370 (D. Rebholz-Schuhmann)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Table 1
Proposed claim-level annotation dimensions

Dimension	Type	Value
Main Claim Text	Free text	Normalised statement of the primary scientific claim of the abstract.
Claim Span	Free text	The sentence(s) from the abstract from which the claim is derived or inferred.
Claim Type	Controlled Vocabulary	The semantic nature of the stated main claim relationship: associative : co-occurrence or correlation asserted without stated mechanism or directionality; causal : a directional mechanistic relationship is stated or linguistically implied; predictive : a prognostic, risk-oriented or biomarker-based assertion; evaluative : a comparative efficacy, safety or functional characterisation between entities.
Claim Polarity	Controlled Vocabulary	The directionality of the finding: positive : the relationship is asserted as present or confirmed; negative : the relationship is absent or refuted; null : no significant association was found.
Assertion strength	Controlled Vocabulary	The degree of epistemic commitment expressed by the authors: committed : a direct factual assertion with no linguistic hedging; hedged : the assertion is qualified by explicit uncertainty markers speculative : the claim is framed as a hypothesis or explicitly requires further study to confirm.
Triple Role	Controlled Vocabulary	The semantic role of each BioRED relation triple relative to the main claim: core claim : the triple directly and specifically instantiates the primary assertion; claim-related : the triple belongs to the same finding but is not the primary assertion (e.g. a secondary outcome or observation); contextual : the triple provides background, methodological context or describes a known relationship not under investigation.

abstracts, creating a human-annotated baseline intended to support future development of claim-aware knowledge representation and automated extraction models.

2. Background and Related Work

Biomedical relation extraction benchmarks, such as BioRED, provide high-quality semantic annotations of entity pairs and their relation types across gene, chemical, variant and disease domains. While BioRED includes a helpful flag to distinguish novel findings from previously established relations, these annotations are restricted to individual, binary triples. They do not group these triples into scientific claims.

Additionally, scientific claim verification benchmarks like SciFact[5] are closer to claim-level representation by pairing atomic claims with abstracts annotated as supporting or refuting them together with rationale sentences. However, SciFact relies on free-text claims synthesized from citation statements rather than structured, multi-dimensional semantic objects extracted directly from the primary research

Table 2

Examples of claim-centered curation from the pilot corpus

Dimension	Annotation Value
Example 1: Contextualizing null results (PMID: 10491763) [11]	
Claim Span	“Mutations in the coding region of the HNF-6 gene are not associated with Type II diabetes or with changes in insulin responses to glucose among the Caucasians examined.”
Main Claim Text	Mutations in the coding region of the HNF-6 gene are not associated with Type II diabetes or altered glucose-induced insulin responses in Caucasian populations.
Categorical Attributes	Type: associative Polarity: null Assertion Strength: committed
Triple Roles	• (Hepatocyte nuclear factor-6, Association, type II diabetes) → <i>core claim</i> • (glucose, Positive_Correlation, insulin) → <i>contextual</i> • (glucose, Positive_Correlation, type II diabetes) → <i>contextual</i>
Example 2: Distinguishing core claims from supporting and contextual relations (PMID: 16720068) [12]	
Claim Span	“This patient presented with symptoms of neuroleptic malignant syndrome (NMS), thus demonstrating that NMS-like symptoms can occur after combined paroxetine and alprazolam treatment. The adverse drug reaction score obtained by the Naranjo algorithm was 6 in our case, indicating a probable relationship between the patient’s NMS-like adverse symptoms and the combined treatment used in this case.”
Main Claim Text	NMS-like symptoms may occur after combined paroxetine and alprazolam treatment in elderly depressive patients.
Categorical Attributes	Type: causal Polarity: positive Assertion Strength: hedged
Triple Roles	• (paroxetine, Association, neuroleptic malignant syndrome) → <i>core claim</i> • (alprazolam, Association, neuroleptic malignant syndrome) → <i>core claim</i> • (agitation, Negative_Correlation, alprazolam) → <i>contextual</i> • (insomnia, Negative_Correlation, paroxetine) → <i>contextual</i>

text. While SciFact demonstrates that claim-level granularity is operationally valuable for downstream inference, it leaves the internal semantic architecture of the claim unencoded.

Understanding what constitutes a well-formed scientific claim requires modeling both its content and the epistemic stance of its authors. Wilbur et al. [6] proposed five orthogonal dimensions for characterising biomedical claims, covering focus, polarity, certainty, directionality and mode of assertion. The BioScope corpus [7] introduced fine-grained annotations for speculation and negation, while FactBank [8] modeled event factuality as a graded epistemic commitment. Together, these resources prove that an author’s degree of commitment and framing can be reliably identified and annotated in scientific text, but they lack a unified schema that binds these disjointed modifiers directly to a set of underlying domain relations to form a structured claim representation.

Vogt and Kuhn [9] propose semantic units as self-contained discourse units that preserve scientific meaning at a granularity finer than full sentences but richer than isolated triples. Similarly, the nanopublication framework [10] provides a standards-compliant metadata infrastructure for wrapping assertions with provenance and publication metadata in named RDF graphs. These frameworks provide the long-term motivation for our target structure to which our pilot annotation serves as a preliminary effort in capturing the kinds of claim-centric properties needed.

3. Annotation Schema and Framework

3.1. Defining Biomedical Main Claims

We define the main claim of a biomedical abstract as the primary novel finding reported by the authors, representing the central scientific contribution of the paper. Main claims differ from background statements, methodological descriptions and secondary findings in that they express the primary result or interpretation supported by the study. Biomedical main claims are most noticeably non-atomic and may emerge through the synthesis of multiple supporting statements distributed across the abstract. We annotate one primary claim per abstract; subordinate findings are captured via Triple Role assignments (See Section 3.3).

3.2. Corpus and Sampling

We compiled a pilot corpus of 150 abstracts sampled from the BioRED training split. To ensure the presence of primary research findings, we restricted the sample to abstracts containing at least one relation annotated as *Novel* within the original BioRED gold standard. Two annotators are independently annotating the corpus. The sample is structurally dominated by *Association* and *Positive_Correlation* labels. This distribution underscores the necessity of our framework: *Association* is the most semantically underspecified relation type in BioRED, making the characterization of its assertion strength and directionality a primary challenge. The MeSH term distribution reflects a broad coverage of study types, with 72% of abstracts involving human subjects, 41% animal models and 9% reporting case-control study designs.

[Home](#) [← Back to list](#) [◀ Prev abstract](#) [Next abstract ▶](#) Annotating as annotator_1

PMID 10491763

#1 out of 150 · done

Hepatocyte nuclear factor-6: associations between genetic variability and type II diabetes and between genetic variability and estimates of insulin secretion.

Abstract

The transcription factor hepatocyte nuclear factor (HNF)-6 is an upstream regulator of several genes involved in the pathogenesis of maturity-onset diabetes of the young. We therefore tested the hypothesis that variability in the HNF-6 gene is associated with subsets of Type II (non-insulin-dependent) diabetes mellitus and estimates of insulin secretion in glucose tolerant subjects. We cloned the coding region as well as the intron-exon boundaries of the HNF-6 gene. We then examined them on genomic DNA in six MODY probands without mutations in the MODY1, MODY3 and MODY4 genes and in 54 patients with late-onset Type II diabetes by combined single strand conformational polymorphism-heteroduplex analysis followed by direct sequencing of identified variants. An identified missense variant was examined in association studies and genotype-phenotype-studies. We identified two silent and one missense (Pro75 Ala) variant. In an association study the allelic frequency of the Pro75Ala polymorphism was 3.2% (95% confidence interval, 1.9-4.5) in 330 patients with Type II diabetes mellitus compared with 4.2% (2.4-6.0) in 238 age-matched glucose tolerant control subjects. Moreover, in studies of 238 middle-aged glucose tolerant subjects, of 226 glucose tolerant offspring of Type II diabetic patients and of 367 young healthy subjects, the carriers of the polymorphism did not differ from non-carriers in glucose induced serum insulin or C-peptide responses. Mutations in the coding region of the HNF-6 gene are not associated with Type II diabetes or with changes in insulin responses to glucose among the Caucasians examined.

+ Add highlighted to Claim Spans

Clear

Selected: "The transcription factor hepatocyte nuclear factor (HNF)-6 is an upstream regula..."

Claim Spans

The transcription factor hepatocyte nuclear factor (HNF)-6 is an upstream regulator of several genes involved in the pathogenesis of maturity-onset diabetes of the young.

Mutations in the coding region of the HNF-6 gene are not associated with Type II diabetes or with changes in insulin responses to glucose among the Caucasians examined.

Figure 1: Claim span selection in the annotation tool. The annotator highlights source sentences from the abstract (PMID 10491763), which are added to the Claim Span field and used to derive the normalized Main Claim Text.

3.3. Annotation Dimensions and Attributes

Our annotation work enriches the underlying data by pairing the annotations inherited from BioRED with claim-level dimensions. The original BioRED attributes including document identifiers, entity types, relation triples and novelty flags are preserved without modification. To these, we apply the new annotation schema summarized in Table 1. For *Main Claim Text*, the annotator writes a normalized, free-text statement of the primary discovery as they understand it, translating complex or fragmented phrasing into a clear summary of the core result. For *Claim Span*, the annotator captures the exact raw sentence(s) from the abstract used to derive or infer that statement. The nature of this assertion is then characterized by its *Claim Type* for which the values are mutually exclusive; *causal* takes priority over *associative* whenever directional or mechanistic language is present. *Assertion Strength* reflects the author’s expressed epistemic commitment within the text itself. For example, a null result can be annotated with a *negative* polarity and a *committed* strength if the authors explicitly state the absence of a relationship. Finally, *Triple Role* maps each co-occurring BioRED relation to the main claim by categorizing it as either a *core claim*, *claim-related* finding or a *contextual* background triple. We illustrate this with two distinct pilot examples in Table 2. Both annotators apply this schema using a purpose-built Streamlit interface (Figures 1 and 2), guided by a shared written annotation guideline. The tool, guideline and resulting corpus will be released publicly upon completion of the pilot.

Claim layer

Main Claim Text

Mutations in the coding region of the HNF-6 gene are not associated with Type II diabetes or with changes in insulin responses to glucose among the Caucasians examined.

Claim Type

Claim Expression Mode

Claim Polarity

Assertion Strength

associative

explicitly stated

negative

committed

Assign Triple Role

Roles: **core claim** (directly instantiates the main claim), **claim-related** (same finding, not the primary assertion), **contextual** (background, methods, comparisons not central).

Entity 1: Hepatocyte nuclear factor-6 Concept ID: 3175 Semantic Type: GeneOrGeneProduct	Relation: Association Novel: No	Entity 2: type II diabetes Concept ID: D003924 Semantic ID: DiseaseOrPhenotypicFeature	Triple Role core claim
Entity 1: glucose Concept ID: D005947 Semantic Type: ChemicalEntity	Relation: Positive_Correlation Novel: No	Entity 2: insulin Concept ID: 3630 Semantic ID: GeneOrGeneProduct	Triple Role contextual
Entity 1: glucose Concept ID: D005947 Semantic Type: ChemicalEntity	Relation: Association Novel: No	Entity 2: type II diabetes Concept ID: D003924 Semantic ID: DiseaseOrPhenotypicFeature	Triple Role contextual

Figure 2: Claim-level annotation interface for the same abstract: Claim Type, Claim Polarity and Assertion Strength for the Main Claim Text, and Triple Role assignment (core claim / claim-related / contextual) for each BioRED relation triple.

4. Conclusion and Future Work

We present an ongoing effort to develop a claim-level annotation schema for biomedical scientific claims grounded in the BioRED corpus. Our schema extends BioRED’s relation annotations with dimensions that capture the type, polarity, epistemic strength and triple-level roles of an abstract’s primary scientific assertion. By organizing these attributes systematically, the schema provides a structured representation of claims that can help make their context and provenance computationally accessible for future knowledge graphs.

The scope of this exploratory work is deliberately restricted to the schema definition and manual annotation phase. We do not model full argumentation structures, discourse relations between separate claims or cross-

document aggregation. Future work will establish formal inter-annotator agreement metrics and refine our guidelines based on the challenges observed during the pilot phase. From there, we plan to scale the annotation framework to the broader BioRED corpus and investigate how these structured claim annotations can support automated extraction models and knowledge representation.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI's GPT-5.5 in order to: Paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] Y. Zhang, X. Sui, F. Pan, K. Yu, K. Li, S. Tian, A. Erdengasileng, Q. Han, W. Wang, J. Wang, J. Wang, D. Sun, H. Chung, J. Zhou, E. Zhou, B. Lee, P. Zhang, X. Qiu, T. Zhao, J. Zhang, A comprehensive large-scale biomedical knowledge graph for AI-powered data-driven biomedical research, *Nature Machine Intelligence* 7 (2025) 602–614. doi:10.1038/s42256-025-01014-w.
- [2] C. Zong, S. Lv, S. tu Xue, H. Zheng, J. Wan, L. Zhang, Building evidence-based knowledge bases from full-text literature for disease-specific biomedical reasoning, 2026. URL: <https://arxiv.org/abs/2603.28325>. arXiv:2603.28325.
- [3] Y. Zheng, W. Liu, B. Zeng, Y. Feng, X. Du, L. Zhou, Y. Li, Automating biomedical knowledge graph construction for context-aware scientific inference, *bioRxiv* (2026). URL: <https://www.biorxiv.org/content/early/2026/01/14/2026.01.14.699420>. doi:10.64898/2026.01.14.699420.
- [4] L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, Z. Lu, Biored: a rich biomedical relation extraction dataset, *Briefings in Bioinformatics* 23 (2022) bbac282. doi:10.1093/bib/bbac282.
- [5] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylén, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, 2020, pp. 7534–7550. URL: <https://aclanthology.org/2020.emnlp-main.609/>. doi:10.18653/v1/2020.emnlp-main.609.
- [6] W. J. Wilbur, A. Rzhetsky, H. Shatkay, New directions in biomedical text annotation: definitions, guidelines and corpus construction, *BMC Bioinformatics* 7 (2006) 356. URL: <https://doi.org/10.1186/1471-2105-7-356>. doi:10.1186/1471-2105-7-356.
- [7] V. Vincze, G. Szarvas, R. Farkas, G. Móra, J. Csirik, The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes, *BMC Bioinformatics* 9 (2008) S9. URL: <https://doi.org/10.1186/1471-2105-9-S11-S9>. doi:10.1186/1471-2105-9-S11-S9.
- [8] R. Sauri, J. Pustejovsky, FactBank: a corpus annotated with event factuality, *Language Resources and Evaluation* 43 (2009) 227–268. URL: <https://doi.org/10.1007/s10579-009-9089-9>. doi:10.1007/s10579-009-9089-9.
- [9] L. Vogt, T. Kuhn, R. Hoehndorf, Semantic units: organizing knowledge graphs into semantically meaningful units of representation, *Journal of Biomedical Semantics* 15 (2024) 7. URL: <https://doi.org/10.1186/s13326-024-00310-5>. doi:10.1186/s13326-024-00310-5.
- [10] P. Groth, A. Gibson, J. Velterop, The anatomy of a nanopublication, *Information services and use* 30 (2010) 51–56.
- [11] A. M. Møller, J. Ek, S. M. Durviaux, S. A. Urhammer, J. O. Clausen, H. Eiberg, T. Hansen, G. G. Rousseau, F. P. Lemaigre, O. Pedersen, Hepatocyte nuclear factor-6: associations between genetic variability and type II diabetes and between genetic variability and estimates of insulin secretion., *Diabetologia* 42 (1999) 1011–1016. doi:10.1007/s001250051261.
- [12] H. Tani, K. Ichihashi, K. Inoue, K. Fujita, Y. Okazaki, Possible neuroleptic malignant syndrome related to concomitant treatment with paroxetine and alprazolam., *Progress in neuro-psychopharmacology & biological psychiatry* 30 (2006) 1176–1178. doi:10.1016/j.pnpbp.2006.04.006.