VaxScope: Document-Level Structured Evidence Extraction from Immunization Systematic Reviews

Bahar İlgen[1], Ebenezer Awotoro[1, 2], Georges Hattab[1,2]
[1] Centre for Artificial Intelligence in Public Health Research (ZKI-PH),
Robert Koch Institute, Berlin, Germany
[2] Department of Mathematics and Computer Science,
Freie Universität Berlin, Berlin, Germany
ilgenb@rki.de

Abstract

Systematic reviews are fundamental to evidence-based medicine, but the clinical evidence they contain is primarily expressed in unstructured text, making large-scale extraction and reuse difficult. Existing biomedical NLP methods have achieved strong performance on span-level extraction from clinical trials and abstracts; however, these approaches are insufficient for systematic reviews, where evidence is often distributed across multiple studies, sentences, and sections and must be aggregated into normalized document-level attributes. We introduce **VaxScope**, a benchmark dataset for document-level structured evidence extraction from immunization-related systematic reviews. VaxScope is constructed through an expert-guided semi-automatic annotation pipeline that combines automatic candidate generation with domain expert validation to ensure consistency and annotation quality. We formalize the task as document-level structured extraction, where target labels are defined at the review level and require aggregating evidence beyond isolated textual spans. We further establish baselines for document-level structured extraction using abstract-level input representations and evaluate how access to evidence-grounded contextual input improves performance over abstract-only settings. Baseline experiments show that PubMedBERT achieves the best overall performance (Avg F1: 0.850), with evidence-grounded input improving performance particularly for fields requiring distributed contextual reasoning.

1 Introduction

Systematic reviews are central to evidence-based medicine, but the evidence they contain is primarily expressed in unstructured natural language, making large-scale extraction, comparison, and reuse difficult for downstream applications such as clinical decision support and evidence synthesis (Pilic et al., 2023). Existing biomedical NLP methods have achieved strong performance on patient-level clinical text, such as electronic health records and clinical notes (Lee et al., 2019; Alsentzer et al., 2019). They have also shown strong results in span-level extraction from biomedical abstracts and trial reports (Fraile Navarro et al., 2023). However, these approaches are often insufficient for systematic reviews, where evidence is synthesized across multiple studies and relevant information is distributed across sentences and document sections rather than localized in isolated spans.

This limitation is particularly important in the immunization domain, where vaccine recommendations and public health policies depend heavily on systematic reviews that summarize evidence across populations, pathogens, interventions, and clinical outcomes (Pilic et al., 2023; İlgen et al., 2024). Supporting decisions in this setting requires structured access not only to explicit entities such as diseases and populations, but also to higher-level attributes such as outcomes, study characteristics, evidence summaries, and review design.

A central challenge is that many of these attributes cannot be reliably extracted from individual text spans alone (Zhang et al., 2024; Schmidt et al., 2025). Instead, they require aggregating evidence across multiple mentions and resolving information at the document level (Zhang et al., 2020; Zheng et al., 2024). For example, the study design or the number of included studies may be mentioned in different sections, while clinically relevant outcomes are often described implicitly across multiple paragraphs. As a result, standard span-based extraction approaches alone are insufficient to capture the full structure of evidence represented in systematic reviews.

To address this limitation, we introduce **VaxScope**, a benchmark dataset for document-level structured evidence extraction from immunization-related systematic reviews. Each review is annotated with a set of normalized attributes, in-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 853–863

cluding population, disease, outcome, study design, number of included studies, and review type. In addition, each structured annotation is linked to a small set of supporting evidence snippets, enabling traceability, human verification, and evidence-grounded extraction and validation. This evidence-grounded design supports interpretable downstream applications and human-in-the-loop decision support (Riskin et al., 2025; DeYoung et al., 2020).

VaxScope is constructed using an expert-guided semi-automatic annotation pipeline that combines automatic candidate generation with domain expert validation to ensure consistency and annotation quality. We establish benchmark baselines for document-level structured extraction using abstract-level input, and demonstrate that multi-task transformer classifiers effectively capture heterogeneous document-level attributes across categorical, multi-label, and numeric fields. These structured attributes can additionally support interactive and visualization-based systems for exploring immunization evidence (İlgen and Hattab, 2025). Taken together, these design choices position VaxScope as both a benchmark dataset and a practical resource for downstream evidence exploration. Our contributions are threefold:

- We introduce **VaxScope**, a new benchmark dataset for document-level structured evidence extraction from immunization-related systematic reviews.
- We design an evidence-grounded annotation schema that combines normalized document-level attributes with linked supporting evidence snippets, enabling traceable, interpretable, and verifiable extraction.
- We establish benchmark baselines for document-level structured extraction using abstract-level input, and evaluate how access to evidence-grounded contextual input improves performance relative to abstract-only settings.

2 Related Work

Biomedical natural language processing has made substantial progress in extracting structured information from clinical and scientific text (Lee et al., 2019). Early approaches relied on rule-based methods using predefined lexical and syntactic patterns, which proved effective for high-precision extraction in constrained settings but often struggled with ambiguous and heterogeneous biomedical language (Wang et al., 2017). More recently, transformer-based models pretrained on biomedical corpora, such as BioBERT (Lee et al., 2019) and BIO_CLINICALBERT (Alsentzer et al., 2019), have substantially improved performance across named entity recognition, relation extraction, and document classification tasks. However, much of this progress has focused on sentence-level or span-level extraction from primary clinical documents, leaving document-level structured evidence extraction from systematic reviews comparatively underexplored.

Multi-label classification of biomedical documents has been studied extensively in the context of literature indexing and topic annotation. Du et al. (2019) proposed ML-Net, a deep learning architecture for multi-label biomedical text classification that jointly predicts labels and label counts. More recently, the BioCreative VII LitCovid track (Chen et al., 2022; Lin et al., 2022) introduced a shared task for multi-label topic annotation of COVID-19 literature, demonstrating the utility of transformer-based classifiers for document-level label assignment in rapidly evolving biomedical domains. While these tasks are closely related to the multi-label components of our framework, they primarily focus on topical document classification. In contrast, VaxScope requires heterogeneous document-level extraction that jointly combines categorical labels, multi-label aggregation, and structured numeric and temporal fields.

A prominent line of work in biomedical NLP focuses on extracting structured evidence from randomized controlled trial (RCT) reports using the PICO framework—Population, Intervention, Comparator, and Outcome. The EBM-NLP corpus (Nye et al., 2018) provided a foundational benchmark for span-level PICO extraction from RCT abstracts and has been widely used to evaluate sequence labeling and NER-based approaches. Subsequent work has explored multi-task learning, section-aware models, and prompt-based methods to improve extraction precision (Hu et al., 2023). More recently, large language models have been applied to PICO extraction in zero-shot and few-shot settings (Ghosh et al., 2024; Chen et al., 2025), achieving competitive performance without requiring labeled training data. However, these approaches remain predominantly span-level and are designed for primary study reports (Lehman et al., 2019). In contrast,

VaxScope targets structured extraction from *systematic reviews*, where evidence is synthesized across multiple primary studies and key attributes often require document-level aggregation rather than isolated span detection.

Automating systematic reviews has attracted growing interest as the volume of biomedical literature continues to expand (Marshall and Wallace, 2019; Tsafnat et al., 2014). Prior work has addressed several stages of the systematic review pipeline, including relevant paper retrieval (van de Schoot et al., 2021), risk of bias assessment (Gates et al., 2018), and evidence synthesis and summarization (Wang et al., 2022). Recent studies have explored the use of LLMs for semi-automated data extraction from systematic reviews, reporting strong performance for extracting predefined entities and study characteristics across multiple domains (Schmidt et al., 2024). At the same time, recent living systematic reviews of systematic review automation show that most existing approaches remain centered on extracting PICO-style evidence from primary studies and randomized controlled trials, while structured extraction of review-level metadata remains comparatively underexplored (Schmidt et al., 2025). In contrast, VaxScope focuses on the review itself as the primary unit of analysis: given a systematic review document, the goal is to extract normalized attributes that characterize the review as a whole, including its scope, methodology, and evidence base. While VaxScope captures methodological attributes such as study design, it does not currently include explicit evidence-quality assessments such as GRADE scores or risk-of-bias ratings. These dimensions fall outside the scope of the current abstract-based benchmarking and typically require a broader full-text methodological appraisal beyond the extraction targets considered here. Incorporating such fields remains an important direction for future dataset extensions.

Standard information extraction systems operate at the sentence or span level, identifying entities and relations within individual sentences. However, in many real-world settings, relevant information is distributed across multiple sentences and sections of a document and cannot be reliably captured through isolated span extraction (Zhang et al., 2020). Document-level relation extraction has emerged as an active research area, with methods that model cross-sentence dependencies through graph neural networks and transformer architectures (Zhou et al., 2021; Yao et al., 2019). In the biomedical domain, document-level extraction has been studied for chemical–disease relation extraction (Li et al., 2016) and other entity-level tasks. VaxScope extends this perspective to structured attribute extraction from systematic reviews, where multiple heterogeneous fields—spanning categorical labels, multi-label sets, and numeric or temporal values—must be jointly extracted at the document level. This heterogeneous formulation distinguishes our task from prior document-level extraction benchmarks, which typically focus on a single relation type or entity category.

3 Methods

3.1 Dataset Construction Pipeline

The VaxScope corpus was constructed through a semi-automatic data creation pipeline designed to balance scalability and annotation quality. We first collected immunization-related systematic reviews and meta-analyses from an expert-curated collection of immunization evidence. For each document, candidate annotations were identified from the full text, as relevant information is often distributed across multiple sections, including methods, results, and discussion. We then applied automatic candidate extraction to identify preliminary values for structured fields such as the number of included studies, date of the last literature search, review type, and disease mentions. For the initial seed set, automatically extracted candidates were subsequently reviewed and corrected through expert validation. This semi-automatic pipeline enabled efficient corpus construction while preserving annotation consistency and clinical relevance, as illustrated in Figure 1.

An existing domain-expert annotated benchmark set of 100 immunization-related systematic reviews served as the foundation for schema design, protocol validation, and final evaluation. We used this benchmark to derive label categories, normalization rules, omission rules, and evidence selection criteria, which were then formalized into a constrained prompt schema for LLM-assisted candidate generation. The model was therefore not used to define labels or annotation logic, but to apply an expert-derived annotation protocol at scale. Candidate generation was performed using `gemini-3-flash-preview` (Google DeepMind), prompted with a constrained JSON schema that defined normalized value spaces, controlled vocab-

ularies, and critical extraction rules for each field (e.g., `date_of_last_lit` was set to `null` if not explicitly stated in the text, to avoid unsupported date assignment). The full prompt schema is provided in Appendix A. Generated candidates were subsequently reviewed and corrected where necessary, while the final evaluation was conducted exclusively on the independently expert-annotated gold set.

Following this expert annotation protocol, the corpus was expanded using an LLM-assisted semi-automatic pipeline to 2,000 annotated documents. During data inspection, we identified 79 duplicate review instances between the expert benchmark and the larger LLM-assisted corpus. To avoid evaluation leakage, these overlapping documents were removed before model development. The remaining corpus consisted of 1,921 documents, which were split into 1,536 training and 385 development instances.

Annotation quality was assessed through two complementary steps: (1) automatic quality evaluation by comparing LLM-generated candidate annotations against the expert gold set, reporting field-level F1 and exact-match accuracy; and (2) inter-annotator agreement analysis on a randomly sampled subset of 50 documents from the expanded corpus. These documents were independently annotated by two of the authors following the expert-derived annotation protocol and without access to the original LLM-generated labels. Agreement was measured using Cohen's κ for categorical fields, Jaccard similarity for multi-label fields, and exact-match accuracy for numeric and temporal fields, as reported in Table 1.

3.2 Task Definition

We formulate the problem as document-level structured evidence extraction from systematic reviews in the immunization domain. Given a review document (typically represented by its abstract), the goal is to map the document to a set of normalized attributes that summarize its key evidence characteristics. Although candidate annotations are identified from the full text during corpus construction, the abstract is used as the primary input for the extraction models, as it provides a consistent and practically deployable representation across studies. This setting differs fundamentally from standard span-based information extraction tasks. In systematic reviews, relevant information is heterogeneous and distributed across multiple sentences and

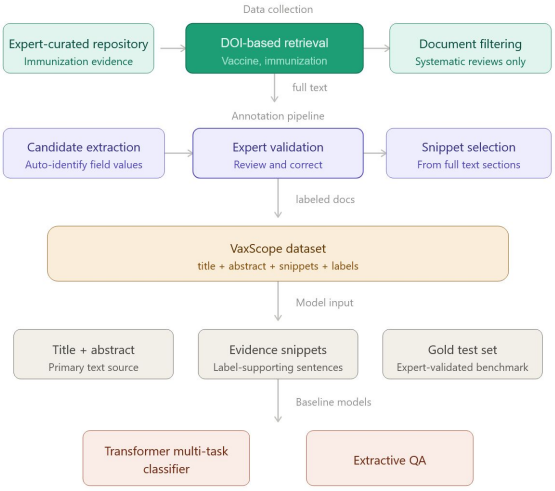

Figure 1: Semi-automatic corpus construction pipeline for VaxScope. Documents are collected from an expert-curated immunization evidence source and retrieved via DOI lookup. Candidate annotations are generated automatically and validated by domain experts. Evidence snippets from the full text provide support for document-level labels.

sections, and cannot be reliably captured through isolated span extraction. For example, population characteristics may appear in the inclusion criteria, outcomes in the results, and study characteristics in the methods. Extracting structured information, therefore, requires consolidating evidence that is often distributed across multiple mentions within a document.

We define the task as predicting a structured representation comprising predefined fields: population, outcome, disease, number of included studies, study design, and review type. These fields correspond to different types of prediction problems: some require normalized categorical classification (e.g., disease, review type), others require aggregating evidence across multiple mentions (e.g., population, study design), and others involve extracting normalized numeric or temporal values (e.g., number of studies, date of last literature search).

To support interpretability and annotation consistency, each document is additionally associated with a small set of evidence snippets (typically a few sentences) that jointly support the extracted attributes, providing traceable justification for document-level annotations and facilitating human verification. In the current benchmark setting, these evidence snippets are included in the annotation schema and used by the baselines as contextual input. However, the current baselines do not ex-

Field	Metric	Score
Disease	Cohen's κ	0.84
Review type	Jaccard	0.89
Num. studies	EM / Near-miss	0.90 / 0.92
Num. participants	EM	0.82
Date of last search	EM / Year-level	0.82 / 0.90
Topic	Jaccard	0.76
Outcome	Jaccard	0.69
Population	Jaccard	0.71

Table 1: Inter-annotator agreement on 50 randomly sampled documents. Cohen's κ is used for disease, Jaccard similarity for multi-label fields, and exact match (EM) with near-miss tolerance for numeric and temporal fields. Near-miss denotes ± 1 for num. studies and year-level match for date of last search. For the outcome field, agreement was computed only over instances where both annotators assigned a non-null value, as null entries reflected topic-conditional field activation rather than outcome-category decisions.

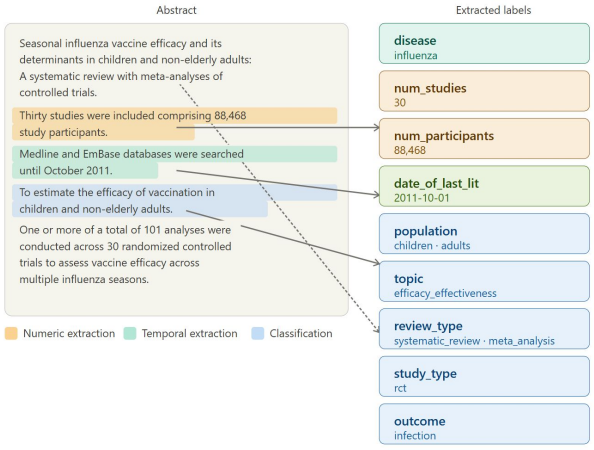

Figure 2: Example annotation from VaxScope. Highlighted spans indicate evidence snippets supporting extracted labels. Amber marks numeric fields, green temporal fields, and blue categorical and multi-label fields. Evidence snippets may be drawn from the abstract or full-text sections depending on where supporting evidence appears in the source document.

plicitly predict snippet selection within their output space.

3.3 Annotation Schema

We define a document-level annotation schema that captures key structured attributes of systematic reviews in the immunization domain. Each document is mapped to a set of predefined fields representing clinically and methodologically relevant aspects of the evidence, such as population, outcome, disease, study design, and review type. Figure 2 shows an annotated example from the VaxScope corpus, illustrating how evidence snippets support the extracted structured labels.

All fields are annotated at the document level, supported by evidence snippets drawn from across the full review. To ensure consistency, each field is associated with a defined value space, including normalized categories (e.g., population groups, study designs) or structured values (e.g., numeric counts and dates). Importantly, the schema is heterogeneous: different fields correspond to different types of prediction problems, including categorical classification, multi-label aggregation over distributed mentions, and numeric or temporal slot filling. This design reflects the diverse structure of information in systematic reviews and aligns with the task formulation described in the previous section. Table 2 summarizes the annotation schema, including field definitions, prediction types, and value spaces. In addition to structured fields, each document is linked to a small set of evidence snippets that provide traceable support for the annotations. The schema is designed to balance expressiveness and the feasibility of annotations. While some fields are normalized into predefined categories, others allow flexible values to capture domain-specific variations. This design supports both structured modeling and practical annotation at scale.

3.4 Baseline Models

We implement a hybrid extraction architecture that combines slot-filling extractors for numeric and temporal fields with a transformer-based multi-task classifier for categorical and multi-label document-level attributes. Figure 3 illustrates the overall extraction architecture.

Multi-task transformer classifier. For categorical and multi-label document-level attributes—including disease, review type, topic, study design, population, and outcome—we evaluate three transformer-based baselines with task-specific prediction heads: BIO_CLINICALBERT and PUBMEDBERT as biomedical encoder baselines, and CLINICAL-LONGFORMER as a long-context baseline. In the primary experimental setting, the model takes the document title and abstract as input and produces a shared document-level representation for downstream field-specific prediction heads. To assess the added value of evidence-aware contextualization, we additionally report an oracle upper-bound setting in which expert-selected evidence snippets are concatenated with the title

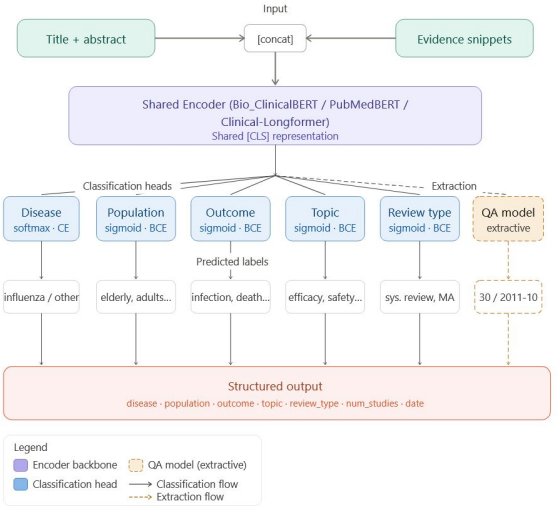

Figure 3: Architecture of the VaxScope extraction system. The shared encoder produces a shared document-level representation that is passed to task-specific classification heads for categorical and multi-label fields. Numeric and temporal fields are extracted separately using an extractive QA model (dashed).

and abstract as additional contextual input. Single-label fields (e.g., disease) are modeled using softmax classification with cross-entropy loss, while multi-label attributes (e.g., population, outcome) are modeled using sigmoid output layers with binary cross-entropy loss.

Numeric and temporal extraction. For fields requiring numeric or temporal values—specifically the number of included studies, the total number of participants, and the date of the last literature search—we adopt an extractive QA approach using `deepset/roberta-base-squad2`, a RoBERTa-based (Liu et al., 2019) question-answering model fine-tuned on SQuAD 2.0 (Rajpurkar et al., 2018). For each target field, we define a natural language question (e.g., *"How many studies were included in this systematic review?"*) and extract the answer span from the document context. Extracted spans are post-processed into typed values (integers or normalized dates in `YYYY-MM-DD` format). Answers below a confidence threshold are mapped to `null`. To validate this design choice, we also compare against a deterministic **regex baseline** that uses regular expressions and domain-specific lexical patterns. This approach can achieve high precision when target values are expressed in predictable surface forms, and serves as a lower-bound reference for the slot-filling fields.

3.5 Evaluation Metrics

Categorical and multi-label fields are evaluated using field-level precision, recall, and F1 score. For numeric slot-filling fields (e.g., number of included studies and total number of participants), we report exact-match accuracy and near-miss accuracy with ± 1 tolerance. This is particularly important in systematic reviews, where multiple related numeric quantities may co-occur within the same document (e.g., numbers screened, excluded, and included), making target disambiguation non-trivial. For temporal fields, we report exact-match accuracy on normalized date strings (`YYYY-MM-DD`) as well as partial-match accuracy at the year level.

4 Experiments

4.1 Experimental Setup

Dataset. VaxScope comprises 1,921 documents drawn from immunization-related systematic reviews, split into 1,536 training and 385 development instances. A held-out set of 100 expert-validated documents is reserved exclusively for final test evaluation.

Models. We evaluate three transformer-based encoders as a shared backbone for the multi-task classifier: BIO_CLINICALBERT (Alsentzer et al., 2019), a biomedical encoder pretrained on clinical notes; PUBMEDBERT (Gu et al., 2021), pretrained exclusively on PubMed abstracts; and CLINICAL-LONGFORMER (Li et al., 2022), a long-context encoder supporting up to 4,096 tokens.

Input settings. Although the labels are defined at the review level and require document-level interpretation, the current baselines use abstract-level input as a standardized and practically accessible benchmark setting. We evaluate two input configurations: a primary setting using only the document title and abstract, and an oracle setting that additionally incorporates expert-selected evidence snippets drawn from the full text. The primary setting reflects a realistic deployment scenario in which only abstract-level information is available. The oracle setting assumes ideal evidence localization and serves as an upper bound on extraction performance. This comparison isolates the contribution of explicit evidence grounding to document-level structured extraction.

Hyperparameters. All models are fine-tuned for 12 epochs using AdamW (lr=2×10^{-5}, weight de-

Field name	Description	Prediction type	Value / label space
Number of studies included	Number of primary studies included in the review	Numeric slot filling	Single integer (e.g., 30)
Date of last literature search	Date of the most recent literature search	Temporal slot filling	Normalized date (stored as YYYY-MM-DD) or null
Review type	Type of review	Multi-label classification	systematic_review, meta_analysis, rapid_review, living_review, umbrella_review, other
Study design	Study designs of included primary studies	Multi-label aggregation	rct, cohort, case_control, cross_sectional, case_series, qualitative, mixed_method, other
Disease	Target disease or pathogen of the review	Single-label classification	Controlled vocabulary (e.g., influenza, covid_19, . . .)
Population	Target population groups	Multi-label aggregation	children, adolescents, adults, elderly, pregnant, healthcare_workers, immunocompromised, all age groups, other
Outcome	Main clinical outcomes for effectiveness analyses	Multi-label classification	infection, hospitalization, death, other
Topic (evidence type)	Primary evidence focus of the review	Multi-label classification	efficacy_effectiveness, immunogenicity, safety, . . .

Table 2: Document-level annotation schema in the VaxScope corpus. All fields are extracted at the document level and supported by selected evidence snippets. Value spaces show representative categories; the full label space is defined in the annotation schema provided with the dataset.

cay=0.01). BIO_CLINICALBERT and PUBMEDBERT use batch size 4 and maximum sequence length 512. CLINICAL-LONGFORMER uses batch size 2 and maximum sequence length 2,048. A sigmoid threshold of 0.5 is applied for multi-label classification.

4.2 Results

Table 3 presents development set results across all model and input setting combinations. We report Macro-F1 for disease classification and Micro-F1 for all multi-label fields.

4.3 Analysis

PUBMEDBERT achieves the best overall performance across both input settings, reaching Avg F1 of 0.797 in the abstract-only setting, 0.810 with evidence snippets on the development set, and 0.850 on the held-out gold test set. This suggests that domain-specific pretraining on PubMed abstracts provides a stronger inductive bias than long-context modeling for this task, likely because systematic review abstracts are structurally dense and domain-specific. CLINICAL-LONGFORMER performs competitively in the abstract-only setting but benefits more from evidence snippets, particularly for *outcome* and *study_type*, suggesting that longer context is most useful when distributed evidence is explicitly provided. Evidence snippets consistently improve performance across models, with

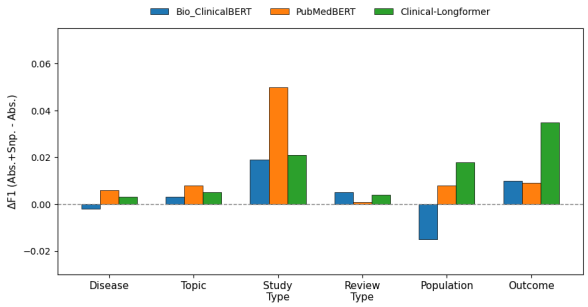

Figure 4: Performance gain from adding expert-selected evidence snippets (Abs.+Snp.) over abstract-only input (Abs.) across prediction fields. Values show ΔF1 on the development set. Gains are largest for study type, population, and outcome, while disease and review type show minimal improvement.

the largest gains observed for *study_type*, *population*, and *outcome*, while *disease* and *review_type* show minimal improvement due to their reliance on standardized labels that are already identifiable from abstracts. The *population* field remains the most challenging across all settings, consistent with inter-annotator agreement results (Jaccard = 0.71), likely reflecting label overlap and implicit population mentions in immunization literature. The *outcome* field also shows relatively lower agreement (Jaccard = 0.69), consistent with its topic-conditional nature.

Model	Setting	Disease	Topic	Study Type	Review Type	Population	Outcome	Avg
Majority baseline	–	0.105	0.324	0.375	0.781	0.284	0.356	0.371
Bio_ClinicalBERT	Abs.	0.945	0.777	0.714	0.948	0.580	0.628	0.765±0.007
Bio_ClinicalBERT	Abs.+Snp.	0.943	0.780	0.733	0.953	0.565	0.638	0.769±0.008
PubMedBERT	Abs.	0.952	0.799	0.752	0.957	0.623	0.698	0.797±0.011
PubMedBERT	Abs.+Snp.	0.958	0.807	0.802	0.958	0.631	0.707	0.810±0.009
Clinical-Longformer	Abs.	0.951	0.785	0.717	0.950	0.591	0.674	0.778±0.009
Clinical-Longformer	Abs.+Snp.	0.954	0.790	0.738	0.954	0.609	0.709	0.792±0.003

Table 3: Development set F1 scores. Macro-F1 is reported for disease, and Micro-F1 for all other fields. Abs. denotes title + abstract only; Abs.+Snp. denotes title + abstract with expert-selected evidence snippets.

Model	Abs.	Abs.+Snp.
Bio_ClinicalBERT	0.806±0.008	0.817±0.009
PubMedBERT	0.838±0.002	0.850±0.006
Clinical-Longformer	0.758±0.001	0.773±0.046

Table 4: Final evaluation results on the held-out expert gold test set. "Abs." uses title + abstract only, while "Abs.+Snp." additionally includes expert-selected evidence snippets. Values report Avg F1, computed as the unweighted mean across task-specific scores (Macro-F1 for disease; Micro-F1 for all remaining fields).

Field	Method	EM	Near-Miss
Num. studies	Regex	0.290	0.290
	QA	0.290	0.290
	QA+Snp.	0.470	0.480
Num. participants	Regex	0.360	0.360
	QA	0.440	0.440
	QA+Snp.	0.600	0.600
Date of last search	Regex	0.020	0.192
	QA	0.172	0.364
	QA+Snp.	0.293	0.576

Table 5: Slot-filling results on the held-out gold test set. EM: exact match; Near-miss: ±1 for numeric fields, year-level for temporal fields. Only documents with non-null gold annotations are evaluated.

5 Discussion

VaxScope addresses a gap in biomedical NLP by providing a benchmark for document-level structured evidence extraction from systematic reviews. Our results show that transformer-based multi-task classifiers can effectively extract heterogeneous structured attributes from immunization-related systematic reviews, achieving strong performance on well-defined fields such as disease and review type. A key finding is that domain-specific pretraining (PUBMEDBERT) is more beneficial than long-context modeling (CLINICAL-LONGFORMER), suggesting that the structural density of systematic review abstracts limits the practical benefit of extended context windows, while familiarity with PubMed-style language provides a stronger inductive bias for this task. Evidence snippets consistently improve performance for fields requiring distributed contextual reasoning, particularly study type, outcome, and population, while highly standardized fields such as disease and review type remain robust under abstract-only input. Lower performance on numeric and temporal slot-filling fields (Table 5) should be interpreted less as model failure and more as a consequence of the intentionally abstract-based benchmark setting, since values such as the number of included studies and the date of the last literature search are often absent from abstracts and only reliably reported in full-text methods sections. In this sense, snippet gains function as indicators of full-text dependency rather than simple performance improvements: large gains for population, outcome, and study type reflect distributed evidence requirements, whereas low-gain fields such as disease and review type are largely recoverable from abstract-only input. Inter-annotator agreement further confirms that well-defined categorical fields such as disease and review type can be annotated consistently, whereas population and outcome remain more challenging due to label ambiguity and implicit mentions. This is also reflected in model performance, where population and outcome remain the most challenging fields across all settings.

6 Conclusion

We introduced VaxScope, a dataset and benchmark for document-level structured evidence extraction from immunization-related systematic reviews. Our baseline experiments show that multi-task transformer classifiers achieve strong performance across heterogeneous document-level fields, with PUBMEDBERT achieving the best overall result (Avg F1: 0.850 on the held-out gold test set). Overall, VaxScope highlights the importance of document-level modeling for structured evidence extraction and provides a foundation for automated evidence synthesis in public health. The VaxScope dataset and annotation schema are available at https://github.com/baharilgen/VaxScope.

Limitations

The current baselines leverage long-context self-attention to capture cross-sentence dependencies implicitly; however, they do not explicitly model mention-level conflict resolution or evidence graphs. More structured aggregation mechanisms, such as hierarchical sentence encoders or graph-based evidence modeling, remain important directions for future work. Furthermore, VaxScope is currently limited to the immunization domain, and generalization to other systematic review domains remains to be explored. In addition, although evidence snippets are included to support interpretability and downstream traceability analyses, the current baselines do not explicitly predict or evaluate snippet selection. Snippet-level grounding and joint label–evidence evaluation also remain important directions for future work.

Acknowledgments

The authors thank the Immunization Unit (FG33) at the Robert Koch Institute for providing the expert-annotated reference set and access to the SYSVAC immunization review registry that supported corpus construction and evaluation.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Qingyu Chen, Alexis Allot, Robert Leaman, Rezarta Islamaj, Jingcheng Du, Li Fang, Kai Wang, Shuo Xu, Yuefu Zhang, Parsa Bagherzadeh, Sabine Bergler, Aakash Bhatnagar, Nidhir Bhavsar, Yung-Chun Chang, Sheng-Jie Lin, Wentai Tang, Hongtong Zhang, Ilija Tavchioski, Senja Pollak, and 20 others. 2022. Multi-label classification for biomedical literature: an overview of the biocreative vii litcovid track for covid-19 literature topic annotations. *Database*, 2022:baac069.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B. Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, Vipina K. Keloth, Kalpana Raja, Jimin Huang, Huan He, Fongci Lin, Jingcheng Du, Rui Zhang, W. Jim Zheng, Ron A. Adelman, and 2 others. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications*, 16(1):3280. Open access.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Jingcheng Du, Qingyu Chen, Yifan Peng, Yang Xiang, Cui Tao, and Zhiyong Lu. 2019. Ml-net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association*, 26(11):1279–1285.

David Fraile Navarro, Kiran Ijaz, Dana Rezazadegan, Hania Rahimi-Ardabili, Mark Dras, Enrico Coiera, and Shlomo Berkovsky. 2023. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, 177:105122.

Allison Gates, Ben Vandermeer, and Lisa Hartling. 2018. Technology-assisted risk of bias assessment in systematic reviews: a prospective cross-sectional evaluation of the robotreviewer machine learning tool. *Journal of Clinical Epidemiology*, 96:54–62.

Madhusudan Ghosh, Shrimon Mukherjee, Asmit Ganguly, Partha Basuchowdhuri, Sudip Kumar Naskar, and Debasis Ganguly. 2024. Alpapico: Extraction of pico frames from clinical trial documents using llms. *Methods*, 226:78–88.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Yan Hu, Vipina K. Keloth, Kalpana Raja, Yong Chen, and Hua Xu. 2023. Towards precise pico extraction from abstracts of randomized controlled trials using a section-specific learning approach. *Bioinformatics*, 39(9):btad542.

Bahar İlgen and Georges Hattab. 2025. Visualizing evidence through natural language interaction for public health interventions. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '24, page 431–438, New York, NY, USA. Association for Computing Machinery.

Bahar İlgen, Antonia Pilic, Thomas Harder, and Georges Hattab. 2024. Pre-training to identify immunization-related entities from systematic reviews. In *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '23, page 234–239, New York, NY, USA. Association for Computing Machinery.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang.

2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *Preprint*, arXiv:2201.11838.

Sheng-Jie Lin, Wen-Chao Yeh, Yu-Wen Chiu, Yung-Chun Chang, Min-Huei Hsu, Yi-Shin Chen, and Wen-Lian Hsu. 2022. A bert-based ensemble learning approach for the biocreative vii challenges: full-text chemical identification and multi-label classification in pubmed articles. *Database*, 2022:baac056.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Iain J. Marshall and Byron C. Wallace. 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8:163.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Antonia Pilic, Sarah Reda, Catherine L Jo, Helen Burchett, Magdalena Bastías, Pauline Campbell, Deepa Gamage, Louise Henaff, Benjamin Kagina, Wiebe Külper-Schiek, Carole Lunny, Melanie Marti, Rudzani Muloiwa, Dawid Pieper, James Thomas, Matthew C Tunis, Zane Younger, Ole Wichmann, and Thomas Harder. 2023. Use of existing systematic reviews for the development of evidence-based vaccination recommendations: Guidance from the sysvac expert panel. *Vaccine*, 41(12):1968–1978.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Daniel Jay Riskin, Keri L. Monda, Joshua J. Gagne, Robert Reynolds, A. Reshad Garan, Nancy Dreyer, Paul Muntner, and Brian D. Bradbury. 2025. Implementing accuracy, completeness, and traceability for data reliability. *JAMA Network Open*, 8(3):e250128.

Lena Schmidt, Kaitlyn Hair, Sergio Graziozi, Fiona Campbell, Claudia Kapp, Alireza Khanteymoori, Dawn Craig, Mark Engelbert, and James Thomas. 2024. Exploring the use of a large language model for data extraction in systematic reviews: a rapid feasibility study. In *Proceedings of the 3rd Workshop on Augmented Intelligence for Technology-Assisted Review Systems (ALTARS 2024)*, Glasgow, UK. CEUR Workshop Proceedings.

Lena Schmidt, Ailbhe N. Finnerty Mutlu, Rebecca Elmore, Babatunde K. Olorisade, James Thomas, and Julian P. T. Higgins. 2025. Data extraction methods for systematic review (semi)automation: Update of a living systematic review. *F1000Research*, 10:401. Originally published 2021-05-19; Version 3 published 2025-04-08.

Guy Tsafnat, Paul Glasziou, Miew Keen Choong, Adam Dunn, Filippo Galgani, and Enrico Coiera. 2014. Systematic review automation technologies. *Systematic Reviews*, 3(1):74. Open access commentary.

Rens van de Schoot, Jonathan de Bruin, Raoul Schram, Parisa Zahedi, Jan de Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, Albert Harkema, Joukje Willemsen, Yongchao Ma, Qixiang Fang, Sybren Hindriks, Lars Tummers, and Daniel L. Oberski. 2021. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3:125–133.

Lucy Lu Wang, Jay DeYoung, and Byron Wallace. 2022. Overview of MSLR2022: A shared task on multi-document summarization for literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 175–180, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2017. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale

document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Gongbo Zhang, Yiliang Zhou, Yan Hu, Hua Xu, Chunhua Weng, and Yifan Peng. 2024. A span-based model for extracting overlapping pico entities from randomized controlled trial publications. *Journal of the American Medical Informatics Association*, 31(5):1163–1171.

Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Tingwen Liu, Hengzhu Tang, Wang Yubin, and Li Guo. 2020. Document-level relation extraction with dual-tier heterogeneous graph. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1630–1641, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hanwen Zheng, Sijia Wang, and Lifu Huang. 2024. A comprehensive survey on document-level information extraction. In *Proceedings of the Workshop on the Future of Event Detection (FuturED)*, pages 58–72, Miami, Florida, USA. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14612–14620.

A Annotation Prompt Schema

The following pseudo-JSON schema was used to prompt `gemini-3-flash-preview` (Google DeepMind) for candidate annotation generation, based on an expert-derived annotation protocol.

```
/* Pseudo-JSON annotation schema */
{
  "doc_id": "doc_XXXX",
  "doi": "...",
  "title": "...",
  "abstract": "...",
  "disease": "...",
  "population": [...],
  "topic": [...],
  "outcome": [...],
  "study_type": [...],
  "review_type": [...],
  "num_studies": null,
  "num_participants": null,
  "date_of_last_lit": null,
  "evidence_snippets": [...]
}
```

A.1 Critical Extraction Constraints

The following constraints were enforced to ensure annotation consistency and reduce schema-inconsistent outputs:

- `date_of_last_lit`:
  set to `null` if not explicitly stated in the text. Publication date must not be used as a substitute.
- `population`:
  if all ages are included or unspecified, use `["all_age_groups"]` alone; no other values may co-occur.
- `disease`:
  use `vaccine_preventable_diseases` only when no single primary disease can be identified.
- `outcome`:
  set to `null` when `topic` does not include `efficacy_effectiveness`; otherwise, one or more of: `infection`, `hospitalization`, `death`, `other`.
- `evidence_snippets`:
  a small set of supporting sentences copied verbatim from the source document; paraphrasing or summarization is not permitted.

A.2 Example Annotation

The following example is drawn from the expert-annotated gold set.

```
{
  "doc_id": "doc_0023",
  "doi": "10.1136/bmjopen-2017-019206",
  "title": "Parents' uptake of human
      papillomavirus vaccines for their
      children: a systematic review and
      meta-analysis of observational
      studies",
  "disease": "human_papillomavirus",
  "population": ["parents_caregivers"],
  "topic": ["coverage", "acceptance"],
  "outcome": null,
  "study_type": ["cross_sectional", "
      cohort", "case_control"],
  "review_type": ["systematic_review", "
      meta_analysis"],
  "num_studies": 79,
  "num_participants": 840838,
  "date_of_last_lit": "2017-11-30",
  "evidence_snippets": [
    "Seventy-nine studies on 840 838
        parents across 15 countries were
         included.",
    "The pooled proportion of parents'
        uptake of HPV vaccines for their
         children was 41.5% (range:
        0.7%-92.8%), twofold higher for
        girls (46.5%) than for boys
        (20.3%).",
    "...",
  ]
}
```