

Validation of a deterministic and reproducible NLP framework for auditable extraction of clinical trial characteristics from PubMed: the MedLitSearch study

Kossi Jean-Jacques Daniel Fianko

kjjdfianko@chu-clermontferrand.fr

Unité de biostatistique et gestion des données, Département de recherche clinique et d'innovation

Nathalie Bourguignon

Unité de biostatistique et gestion des données, Département de recherche clinique et d'innovation

Aurélié Cabrespine

Unité de biostatistique et gestion des données, Département de recherche clinique et d'innovation

Sylvain Mathieu

Centre Hospitalier Universitaire de Clermont-Ferrand

Florian Martinet-Kosinski

Centre Hospitalier Universitaire de Clermont-Ferrand

Bruno Pereira

Unité de biostatistique et gestion des données, Département de recherche clinique et d'innovation

Research Article

Keywords: PubMed, Natural Language Processing, Information Extraction, Text Mining, Reproducibility of Results, Systematic Reviews, Evidence-Based Medicine

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11173070/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background.

Modern evidence synthesis workflows require information extraction methods that are reproducible, traceable and auditable. Generative approaches may accelerate some steps of literature searching, but their output variability and limited auditability may be problematic when extracted data must be compared with an expert standard or reused in methodological analyses. This study aimed to validate the reproducibility and auditability of MedLitSearch, a deterministic framework for extracting key clinical trial characteristics from PubMed.

Methods.

The strategy was operationalized in MedLitSearch, a pipeline interfaced with the NCBI E-utilities API. It combines batch PubMed retrieval and three deterministic natural language processing (NLP) modules evaluated in this study: Outcome Extractor, Participant Extractor and Region Detector. Validation included three independent levels: 251 automated PubMed retrieval tests; an automatic benchmark on 200 PubMed abstracts (N1 corpus); and blinded double expert annotation of 100 independent articles (N2 corpus), stratified across 10 clinical or methodological categories. Pre-specified thresholds were F1 $\geq 78\%$, precision $\geq 80\%$, recall $\geq 75\%$, MAE < 50 on a homogeneous corpus for sample size, and region accuracy $\geq 85\%$.

Results.

PubMed retrieval passed 251/251 tests (100%). In N1, Outcome Extractor achieved 98.2% precision, 99.1% recall and an F1-score of 98.7%. Participant Extractor achieved an F1-score of 92.4% and a MAE of 16.2 participants among abstracts with an extractable sample size. Region Detector achieved 100.0% accuracy on 122 usable affiliations. In N2, inter-annotator agreement was substantial for the primary outcome ($\kappa = 0.761$) and almost perfect for region ($\kappa = 0.966$). Outcome Extractor reached F1 = 100.0% against the expert consensus. Participant Extractor remained robust against consensus (F1 = 88.0%), with numerical error influenced by observational or fragmented cohorts.

Conclusions.

This study proposes and validates a deterministic framework for auditable extraction of clinical trial characteristics from PubMed. The results support deterministic NLP approaches as a reproducible layer to assist evidence synthesis, provided that human oversight and context-specific validation are maintained. Multilingual generalization, time savings and prospective validation of the multicenter detector require dedicated evaluation.

1. Introduction

Biomedical evidence synthesis relies on bibliographic search and data extraction procedures that must be reproducible, transparent and sufficiently structured to allow comparison and audit of results [1]. PubMed is a major source of biomedical literature, yet repeated identification and extraction of methodological characteristics of clinical trials remain difficult when information must be collected article by article.

In systematic reviews and clinical meta-research, the challenge is not only to retrieve relevant references. It is also to extract variables such as the primary outcome, sample size and geographical origin of studies in a consistent manner. Existing bibliographic search systems show variable performance for systematic review needs [2]. At the same time, generative artificial intelligence tools and review-assistance systems may accelerate certain steps, but they raise methodological concerns regarding output stability, traceability of extracted values and auditability of decisions [3–5].

The variables targeted by MedLitSearch were selected because they represent complementary dimensions that are frequently useful when structuring clinical trial evidence. The primary outcome describes the evaluative target of the trial; sample size informs the scale and potential weight of the study; geographical region provides contextual information derived from affiliations; and multicenter status, although not included in the main validation results, reflects study organization and potential external validity.

Natural language processing applied to biomedical texts has been extensively studied, including through PICO and EBM-NLP corpora [6]. These studies have shown that automatic extraction of methodological elements from the clinical literature is difficult because of heterogeneous wording, reporting ambiguities and possible disagreement between human annotators. In comparable clinical and biomedical annotation studies, moving from an automatic or semi-structured standard to a human gold standard is often associated with a decrease in performance, sometimes in the range of 10 to 20 percentage points, because of borderline cases identified by human annotation [10, 11].

In this context, deterministic rule-based approaches should not be understood as a general alternative to generative models, but as a methodological response to a specific need: producing stable, inspectable and directly verifiable extraction for variables defined a priori. This distinction is important in evidence synthesis workflows, where the same input should yield the same output, extraction rules should be auditable, and extracted values should remain linked to the source text. This distinction is important in evidence synthesis workflows, where the same input should yield the same output, extraction rules should be auditable, and extracted values should remain linked to the source text.

The objective of this study was to validate a deterministic methodological strategy for reproducible and auditable extraction of key clinical trial characteristics from PubMed. This strategy was operationalized in MedLitSearch and validated using a multi-level framework combining technical validation of bibliographic retrieval, an automatic benchmark and independent double expert annotation. The study

therefore does not merely describe a software tool; rather, it tests a validation framework applicable to extraction systems intended for evidence synthesis and clinical meta-research.

2. Methods

2.1 Study design and level nomenclature

This methodological study evaluated a deterministic NLP framework for extracting key clinical trial characteristics from PubMed. MedLitSearch is the operational implementation of this framework. The evaluation was structured into three global levels: (i) technical validation of PubMed retrieval; (ii) automatic benchmarking on the N1 corpus; and (iii) human validation through double expert annotation on the N2 corpus. These three levels addressed distinct validation questions: retrieval reliability, technical agreement with an automatic reference, and robustness to human interpretation. To avoid ambiguity, the internal numbering of PubMed retrieval validation tests (four test families detailed in Appendix B and Supplementary Table S2) is distinct from the N1 and N2 designations, which refer exclusively to the two NLP evaluation corpora. The N1 and N2 corpora were built independently to limit the risk of data leakage across validation levels.

2.2 Brief description of the pipeline

The pipeline is directly interfaced with PubMed through the NCBI Entrez Programming Utilities [9]. PubMed was used as the sole data source. Implementation-specific details, including API parameters, software environment, exports and technical reproducibility constraints, are reported in Appendix E to keep the main manuscript focused on the methodological design and validation strategy.

The choice of a deterministic rule-based framework was part of the methodological design rather than a purely technical implementation decision. It was motivated by the need for stable outputs, inspectable rules and direct traceability of extracted values to the source text. The methodological positioning of this deterministic approach with respect to generative or probabilistic approaches is summarised in Supplementary Table S1.

2.3 Evaluated modules

Outcome Extractor. identification of the primary outcome from direct and inferential markers in abstracts.

Participant Extractor. extraction of sample size or total number of participants when this information is explicitly available.

Region Detector. assignment of a broad geographical region from the usable affiliation of the last author.

Multicenter Detector. a developed and integrated module, but not included in the main validation results; its prospective validation is presented as a future direction.

2.4 Metrics, definitions and pre-specified thresholds

Classification performance was described using confusion-matrix components: true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN). Precision was defined as $TP/(TP + FP)$, recall as $TP/(TP + FN)$, and the F1-score as the harmonic mean of precision and recall [7, 8]. The F1-score was retained as the main metric for Outcome Extractor and Participant Extractor because it summarizes the trade-off between extraction errors and omissions in a setting where classes may be imbalanced, consistent with evaluation practices in biomedical NLP [6, 8].

For Participant Extractor, when the expected output was numerical, mean absolute error (MAE) was calculated as the mean absolute difference between the extracted sample size and the reference sample size [7]. For Region Detector, accuracy was defined as the proportion of usable affiliations for which the region assigned by the system matched the reference region. In N1, this reference region was generated automatically from the explicit country detected in the last-author affiliation string and then mapped to a harmonized geographical taxonomy. In N2, the reference region was provided by experts, who applied the same last-author affiliation rule. Thus, region accuracy in N1 measures agreement with an automatic affiliation-based reference, whereas region accuracy in N2 measures agreement with human annotation.

Validation thresholds were defined a priori, before analysis, to limit the risk of post hoc interpretation. They were selected to reflect both common evaluation practices in biomedical information extraction and operational requirements for use in institutional clinical research. Confidence intervals were not calculated because the validation was designed as a descriptive assessment on fixed corpora rather than inferential sampling from a target population; uncertainty was addressed through stratified reporting, effective denominators and inter-annotator agreement.

The pre-specified metrics and validation thresholds are summarised in Table 1.

Table 1
Pre-specified metrics and thresholds.

Module / task	Main metric	Pre-specified threshold
Outcome Extractor	Precision, recall, F1-score	$P \geq 80\%$, $R \geq 75\%$, $F1 \geq 78\%$
Participant Extractor – detection	Precision, recall, F1-score	$P \geq 80\%$, $R \geq 75\%$, $F1 \geq 78\%$
Participant Extractor – numerical value	MAE on homogeneous corpus	$MAE < 50$
Region Detector	Accuracy	$Accuracy \geq 85\%$

2.5 Level 1: PubMed retrieval validation

PubMed retrieval was evaluated using 251 automated tests grouped into four families: offline unit tests, targeted PMID retrieval, keyword search scenarios and retrieval of specific known articles. These tests

aimed to verify query construction, journal filters, pagination and retrieval of metadata required by the pipeline. The number of 251 tests was not based on a statistical sample-size calculation; it corresponded to the complete set of automated checks available at the time of validation, covering the main functional retrieval scenarios required by the pipeline. Details of these test families are provided in Supplementary Table S2.

2.6 Level 2: automatic NLP benchmark on the N1 corpus

The N1 corpus included 200 PubMed abstracts. This number was determined pragmatically: it represented an increase from the initial 150-abstract pilot to reduce sampling variability while keeping the benchmark inspectable, reproducible and compatible with automated execution. N1 was not intended to provide a definitive estimate of generalizability, but to test the technical behavior of the extraction rules under controlled automatic-reference conditions.

2.6.1 Definition of “Predicted” and “Actual” in N1

In the N1 benchmark, both “Predicted” and “Actual” values were generated automatically, without human annotation. The “Predicted” value corresponded to the direct output of the evaluated NLP module — Outcome Extractor, Participant Extractor or Region Detector — applied to each PubMed abstract retrieved through the NCBI E-utilities API. The “Actual” value corresponded to an independent heuristic reference implemented separately in the benchmark script and applied to the same raw text.

This heuristic reference searched for explicit lexical markers: for example, combinations of “primary” or “main” with “outcome(s)” or “endpoint(s)” for Outcome Extractor; sample-size expressions associated with terms such as “enrolled” or “randomized” for Participant Extractor; and explicit country names in affiliation strings for Region Detector. This reference was designed to be strict and conservative in order to limit false positives, but it was not verified by human reviewers.

Because the heuristic reference and the evaluated modules partly rely on overlapping lexical evidence, agreement measured in N1 should not be interpreted as a fully independent validation of the tested system. N1 should be understood as a technical benchmark assessing the consistency of the extraction framework with an automatically generated reference under explicit reporting conditions. Independent validation against human experts was performed separately in N2.

For Outcome Extractor, “Predicted positive” meant that the module had extracted a non-empty primary outcome; “Actual positive” meant that the heuristic reference explicitly detected the presence of a primary-outcome marker. For Participant Extractor, evaluation assessed detection of an extractable sample size and, when a unique value was available, numerical error. For Region Detector, the region predicted by the module was compared with the reference region automatically derived from the explicit country in the last-author affiliation.

2.7 Level 3: double expert annotation on the N2 corpus

The N2 corpus included 100 articles independent from N1, stratified into ten categories: cardiac surgery, liver surgery, pediatric surgery, oncological surgery, case series, meta-analyses, neurosurgery, paramedical care, observational studies and orthopedics. The sample size of 100 articles was a pragmatic feasibility choice for manual double expert annotation: 10 clinical or methodological categories, with 10 articles per category. This design aimed to assess robustness across heterogeneous contexts and identify major extraction failures, rather than to provide a definitive estimate of generalizability across all PubMed-indexed clinical trials.

Two experts annotated the articles independently and blinded to each other's assessments. The experts were not named in the reporting protocol; their profiles were described as follows: one clinician involved in surgical clinical research and one methodologist experienced in clinical research. Their role was to provide a human reference for the evaluated variables, without access to the other expert's assessment during initial annotation.

Inter-annotator agreement was estimated using Cohen's kappa on cases usable for each task. System outputs were compared separately with Expert 1, Expert 2 and the expert consensus. Individual comparisons assessed agreement with each annotator. Consensus-based evaluation was restricted to cases for which both experts provided concordant annotations. Disagreement cases were retained for expert-specific comparisons but excluded from consensus-based performance estimates. No third annotator was used to adjudicate disagreements.

2.8 Comparison rules between system outputs and reference standards

For Outcome Extractor in N2, the reported evaluation assessed binary detection of the presence or absence of an explicitly mentioned primary outcome. The F1-score therefore does not measure textual or semantic accuracy of the extracted content, but agreement between the system and the experts on whether an identifiable primary outcome existed. Text fields annotated by the experts were used for qualitative purposes, but not for F1-score or kappa calculation. This distinction is important: different textual formulations could be counted as agreement if both the system and the expert concluded that an explicit primary outcome was present.

For Participant Extractor, evaluation distinguished detection of an extractable sample size from numerical agreement. When several competing values were reported — by arm, subpopulation, source cohort, registry or analyzed population — the article was considered potentially ambiguous. System-reference pairs were included in the MAE calculation only when a unique value could be matched. Other cases contributed to qualitative error interpretation and explain the difference between the total number of N2 articles and the number of pairs used for stratified numerical analysis.

For Region Detector, the evaluated variable was a broad geographical region derived from the last-author affiliation, not the actual country of trial recruitment or conduct. The same last-author affiliation rule was applied to the system and to the experts. This choice relies on the approximation that the last author

often represents the coordinating or supervisory site of the study. This approximation remains imperfect, particularly for multicenter or international studies, or when affiliations do not reflect recruitment sites.

3. Results

3.1 PubMed retrieval validation

All 251 automated PubMed retrieval tests were successful (251/251; 100%). This confirms the reliability of the retrieval chain in the tested scenarios, including PMID retrieval, keyword searches and verification of journal filters. The detailed organization of these tests is presented in Supplementary Table S2.

3.2 N1 performance

Outcome Extractor was evaluated on the 200 abstracts of the N1 corpus. According to the automatic standard, 112 abstracts contained an identifiable primary outcome and 88 did not. The module correctly identified 111 primary outcomes, with 1 false negative and 2 false positives, corresponding to 98.2% precision, 99.1% recall and an F1-score of 98.7%.

The N1 confusion matrix for Outcome Extractor is shown in Table 2.

Table 2
Confusion matrix – Outcome Extractor (N1, n = 200 abstracts).

Automatic reference: present	Predicted: primary outcome present	Predicted: primary outcome absent
	TP = 111	FN = 1
Automatic reference: absent	FP = 2	TN = 86

In this matrix, “Predicted” refers to the output of Outcome Extractor and “Automatic reference” to the “Actual” value produced by the heuristic defined in the Methods section. The matrix therefore describes agreement between two automatic procedures, rather than agreement with human annotation.

The three residual errors corresponded to two false positives caused by quasi-protocol wording without an explicit primary objective, and one false negative linked to an indirect comparative formulation of the primary outcome. This false negative belonged to the N1 corpus and did not appear in N2, consistent with the PMID-based independence check.

Participant Extractor was evaluated on abstracts containing an identifiable sample size. The module achieved an F1-score of 92.4%, with 86.6% precision, 99.0% recall and a MAE of 16.2 participants, below the pre-specified threshold (MAE < 50) on this homogeneous corpus. The largest errors involved fragmented sample sizes or subpopulations when several numerical values coexisted in the abstract. The largest absolute errors observed in N1 are presented in Supplementary Table S3.

Region Detector was evaluated on 122 usable affiliations. The module achieved 100.0% accuracy, with no observed errors in this corpus. The detailed geographical distribution of N1 is presented in Supplementary Table S4.

3.3 N2 performance against expert annotation

The N2 corpus included 100 independent articles stratified across 10 categories. Inter-annotator agreement was substantial for primary outcome identification ($\kappa = 0.761$) and almost perfect for region ($\kappa = 0.966$). Crude agreement was 87.9% for the primary outcome and 97.8% for region; these proportions describe observed agreement and should not be interpreted as equivalent to kappa, which accounts for agreement expected by chance. The analyzed denominators differed slightly across reference standards and tasks because of unusable information or disagreement cases excluded from consensus-based analyses. Performance was calculated separately against Expert 1, Expert 2 and consensus, according to the rules described in the Methods.

3.3.1 Outcome Extractor

Against Expert 1, Outcome Extractor achieved an F1-score of 91.3% (precision = 100.0%, recall = 83.9%). Against Expert 2, it achieved an F1-score of 96.7% (precision = 93.6%, recall = 100.0%). Against the expert consensus, the F1-score was 100.0% (precision = 100.0%, recall = 100.0%). These results should be interpreted as binary detection of primary-outcome presence, not as a measure of textual correspondence of the extracted label.

Category-level N2 results for Outcome Extractor against Expert 1 are shown in Table 3.

Table 3
Outcome Extractor performance by category (N2, comparison against Expert 1). Precision = 100.0%; recall = 83.9%.

Category	TP	FP	FN
Cardiac surgery	10	0	2
Case series	8	0	0
Liver surgery	10	0	0
Meta-analyses	6	0	2
Neurosurgery	9	0	0
Paramedical care	8	0	0
Observational studies	5	0	2
Orthopedics	5	0	2
Pediatric surgery	10	0	0
Oncological surgery	10	0	3
Total	81	0	11

3.3.2 Participant Extractor

Against the human consensus, Participant Extractor achieved an F1-score of 88.0% (precision = 82.5%, recall = 94.3%). The two experts agreed on sample size in 86.2% of cases. Numerical analysis showed that absolute errors were strongly influenced by observational studies or abstracts describing several subsample sizes. The consensus-based MAE reached 6,969.1 participants and was driven by extreme or ambiguous sample-size values, particularly in categories involving large cohorts or multiple competing numerical values. MAE observed in expert-specific comparisons was higher than the consensus-based MAE, suggesting that extreme or ambiguous sample-size values had a stronger influence when individual expert annotations were considered. By contrast, consensus-based evaluation was restricted to cases in which both experts provided concordant annotations, thereby excluding some disputed or difficult-to-match values. MAE should therefore be interpreted together with category stratification rather than as a homogeneous measure applicable to all study types.

Category-level N2 results for Participant Extractor are shown in Table 4.

Table 4
Participant Extractor performance by category (N2).

Category	n pairs	F1 (%)	MAE	Interpretation
Cardiac surgery	8	75.0	54	Registries or competing values
Case series	4	75.0	19	Small sample sizes generally well detected
Liver surgery	9	88.9	6	Explicit sample sizes
Meta-analyses	7	100.0	0	Clearly reported total sample sizes
Neurosurgery	4	75.0	2,083	Cohorts or multiple arms
Paramedical care	9	88.9	18	Generally explicit sample sizes
Observational studies	3	66.7	51,234	Large historical cohorts
Orthopedics	14	82.4	2,286	Subsamples and multiple cohorts
Pediatric surgery	5	80.0	101	Moderate sample sizes
Oncological surgery	9	88.9	13	Explicit sample sizes

Overall, 72 pairs were matchable for stratified sample-size analysis. The difference from the 100 N2 articles is explained by the absence of usable sample-size information, annotation disagreements, or the impossibility of matching a unique value between the system and the reference.

3.3.3 Region Detector

Against Expert 1, Region Detector achieved 90.7% accuracy (88/97). Against Expert 2, accuracy was 94.8% (91/96). Against the consensus, it reached 91.8% (90/98). Denominators varied slightly (97, 96 and 98) according to the reference standard because articles without a usable last-author affiliation or classified as disagreements for the consensus comparison were excluded from the corresponding denominator. Residual errors were mainly related to absent or truncated affiliations, multinational affiliations, or discordance between the last-author affiliation and the actual place where the study was conducted. The detailed geographical distribution of N2 affiliations is provided in Supplementary Table S5.

3.4 Cross-level synthesis

N1 and N2 do not constitute a test–retest on the same corpus or the same reference standard. N1 measures agreement with a strict automatic reference based on explicit markers, whereas N2 measures agreement with independent human experts. Comparisons between N1 and N2 should therefore be interpreted as a cross-level consistency check rather than as a direct estimate of performance loss on an identical task. The N1 benchmark is useful for testing technical consistency, but its reference standard is not fully independent from the deterministic logic of the modules; N2 provides the main independent human validation layer.

The cross-level synthesis of module performance is presented in Table 5.

Table 5
Performance synthesis by level and reference standard.

Module	Metric	N1	N2 Expert 1	N2 Expert 2	N2 consensus
Outcome Extractor	Precision	98.2%	100.0%	93.6%	100.0%
	Recall	99.1%	83.9%	100.0%	100.0%
	F1-score	98.7%	91.3%	96.7%	100.0%
Participant Extractor	Precision	86.6%	73.6%	79.1%	82.5%
	Recall	99.0%	94.4%	97.1%	94.3%
	F1-score	92.4%	82.7%	87.2%	88.0%
	MAE	16.2	—	—	6,969.1 (see 4.4)
Region Detector	Accuracy	100.0%	90.7%	94.8%	91.8%

Compared with Expert 1, Outcome Extractor's F1-score decreased from 98.7% in N1 to 91.3% in N2, a difference of – 7.4 points. This magnitude remains below decreases sometimes reported in comparable clinical and biomedical annotation studies when moving to a human standard, often in the range of 10 to 20 percentage points [10, 11]. Compared with the consensus, the F1-score reached 100.0% in N2; this value should not be directly combined with the comparison against Expert 1 because the evaluation reference standards differ. No confidence interval was estimated for this descriptive fixed-corpus comparison.

4. Discussion

4.1 Main finding

This study shows that a deterministic NLP strategy can combine batch PubMed retrieval and traceable extraction of key clinical trial characteristics, provided that it is evaluated through a multi-level validation strategy. The main finding is not only the numerical performance of MedLitSearch, but the demonstration that a deterministic extraction layer can be validated as a reproducible and auditable method for variables defined a priori.

4.2 Methodological contribution

The methodological contribution lies in the integration of three requirements that are rarely evaluated together: reproducibility of bibliographic retrieval, determinism of extraction and comparison with an expert standard. This approach addresses a need in systematic reviews and clinical meta-research, where extraction decisions must be auditable, repeatable and documented. MedLitSearch should

therefore be interpreted as the implementation of a methodological framework rather than as a mere software demonstration.

4.3 Complementarity between N1 and N2

A central methodological point is the complementarity between N1 and N2. N1 provides a reproducible automatic benchmark for verifying the technical behavior of deterministic extraction rules when explicit textual markers are present. However, because the N1 reference is heuristic and partly based on lexical evidence that overlaps with the modules, it cannot be considered a fully independent gold standard. N2 addresses this limitation by comparing system outputs with independent blinded human expert annotations. This organization distinguishes technical consistency from robustness to expert interpretation.

4.4 Interpretation of NLP performance

Outcome Extractor performance was high across both validation levels. The consensus result should nevertheless be interpreted cautiously: it does not mean that the task is solved in all contexts, but rather that, in this corpus, remaining disagreements corresponded to cases of human ambiguity. In addition, the F1-score reported for Outcome Extractor measures binary detection of the presence of an explicit primary outcome, not exact textual correspondence of the extracted content.

Participant Extractor performance highlights a classic limitation of numerical extraction: abstracts may mention sample sizes by arm, subpopulations, registries, source cohorts or analyzed populations that differ from the included population. The overall MAE against consensus (6,969.1 participants) illustrates this phenomenon: it was almost entirely attributable to observational studies (MAE = 51,234) and neurosurgery (MAE = 2,083), where abstracts described subpopulations from very large cohorts. After excluding these two categories, the median of category-level MAEs remained below 140 participants. This limitation supports future disambiguation rules for multiple sample sizes and dedicated validation on corpora enriched with large cohort studies.

Region Detector maintained accuracy above the pre-specified threshold against human standards. Observed errors mainly reflected the quality of PubMed metadata, including missing, incomplete or non-representative affiliations. Because region is derived from the last-author affiliation, it should be interpreted as an affiliation-based contextual indicator rather than as a measure of the actual recruitment or conduct location of the trial.

4.5 Positioning with respect to generative approaches

This study does not aim to establish a general superiority of deterministic systems over generative approaches. Large language models may be useful for exploration, reading assistance, summarization or hypothesis generation. However, when the objective is to extract methodological variables defined a priori in a workflow that must be audited, output stability and rule traceability become central. In this context, a deterministic approach offers a methodological advantage: it enables direct, repeatable and versionable validation of outputs against an automatic or human standard.

The main risk of generative approaches in this setting is not only extraction error, but also the difficulty of distinguishing a value that was extracted from one that was reformulated, inferred or not directly sourced. Conversely, a deterministic system may still make mistakes, particularly in the presence of multiple numbers or ambiguous affiliations, but the error is generally inspectable and attributable to a rule or identifiable textual pattern. This distinction supports deterministic approaches as an initial extraction layer in evidence synthesis workflows.

4.6 Implications for evidence synthesis and meta-research

Beyond the performance of the evaluated modules, this study proposes a validation framework applicable to extraction systems intended for evidence synthesis. This framework combines technical validation of bibliographic retrieval, a controlled automatic benchmark, independent double expert annotation, comparison with individual and consensus reference standards, and pre-specified performance thresholds. This organization helps distinguish technical performance, robustness to human ambiguity and the actual scope of the generated outputs.

This type of approach may support several meta-research uses, including surveillance of clinical trial landscapes, initial structuring of corpora for systematic reviews, methodological trend analyses, monitoring of reported outcomes and rapid identification of studies requiring priority human verification. These uses should nevertheless be considered as support for structuring and audit, not as a replacement for methodological expertise.

4.7 Comparison with the literature and caution regarding novelty

The results are consistent with the biomedical NLP literature applied to clinical trials, where PICO and EBM-NLP corpora have shown the difficulty of automatically extracting Population, Intervention, Comparator and Outcome elements [6]. The observed decrease between automatic and human standards (– 7.4 points for Outcome Extractor against Expert 1) remains below differences sometimes reported in comparable clinical and biomedical annotation studies [10, 11]. To our knowledge, few published frameworks jointly document batch PubMed retrieval validation and deterministic NLP extraction evaluated against an expert standard. This statement should be understood as cautious positioning, not as a claim of exhaustive uniqueness in the absence of a formal systematic review of competing tools.

4.8 Practical implications

In conditions close to institutional systematic review workflows, a deterministic pipeline could support prescreening, initial data structuring and extraction audit. However, time savings compared with manual extraction were not formally measured in this study. They should be considered an operational hypothesis to be confirmed through prospective evaluation, for example using a user study and a standardized usability measure such as the System Usability Scale [12–14].

4.9 Limitations

- The N1 corpus relies on an automatic heuristic reference based on explicit markers; it may overestimate performance on atypical formulations.
- The N1 reference was not verified by human annotators and is not fully independent from the deterministic rules evaluated, because of partial overlap in lexical evidence.
- The N2 corpus is limited to 100 articles and 10 categories. This pragmatic expert-annotation sample was sufficient to assess concordance and identify major failures, but it is not a definitive estimate of generalizability.
- Disagreements between experts were not adjudicated by a third annotator; disputed cases were excluded from consensus-based evaluation rather than resolved.
- NLP extraction is primarily adapted to English and French abstracts; multilingual coverage remains partial.
- Participant Extractor is sensitive to fragmented sample sizes, subpopulations and large observational cohorts.
- Multicenter Detector has been developed but has not yet been validated on an annotated gold standard.

5. Conclusion

This study proposes and validates a deterministic NLP framework for reproducible and auditable extraction of key clinical trial characteristics from PubMed. MedLitSearch operationalizes this framework by combining batch PubMed retrieval, structured extraction and multi-level validation against automatic and human standards. The results support deterministic approaches when reproducibility, traceability and auditability are priorities. Conclusions should nevertheless remain limited to the evaluated corpora: N1 primarily documents technical consistency with a strict heuristic reference, whereas N2 provides independent validation against human experts. Next steps should focus on prospective validation of Multicenter Detector, improved disambiguation of multiple sample sizes, multilingual extension and formal evaluation of time savings and usability in real-world settings.

Declarations

Ethics approval and consent to participate. Not applicable. This methodological study used PubMed bibliographic records, abstracts and published scientific articles. It did not involve human participants, individual patient-level data or identifiable personal data. No ethics committee approval was required according to the nature of the data used.

Consent for publication. Not applicable.

Clinical trial number: Not applicable.

Funding. This study received no specific funding.

Competing interests. The authors declare no competing interests.

Availability of data and materials. The validation datasets supporting the findings of this study are available on Zenodo at: <https://doi.org/10.5281/zenodo.21873453>. The archive contains derived validation materials and metadata required to reproduce the reported results, including N1 benchmark outputs, N2 article list and categories, anonymized expert annotations, consensus indicators, system outputs, scoring results and source tables. Full-text articles and publisher-owned content are not redistributed.

Code availability. The MedLitSearch source code is available on GitHub under the MIT License. The code version corresponding to the analyses reported in this manuscript was identified by commit d860cf1e6336d627b89054593f06842693f0946d.

Authors' contributions. KJJDF conceived the study, developed the pipeline, curated the datasets, conducted the analyses and drafted the manuscript. NB contributed to methodological design, interpretation of the validation strategy and critical revision. AC contributed to methodological review. SM contributed clinical and methodological expertise and critical revision. FMK contributed clinical interpretation and critical revision. BP contributed methodological supervision, statistical interpretation and critical revision. All authors read and approved the final manuscript.

Acknowledgements. The authors thank the MAESTRO group for methodological discussions and scientific support. The MAESTRO group did not provide specific funding for this study.

References

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ*. 1996;312(7023):71–2.
2. Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Res Synth Methods*. 2020;11(2):181–217. 10.1002/jrsm.1378.
3. Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*. 2021;372:n160. 10.1136/bmj. n160.
4. Lau O, Golder S. Comparison of Elicit AI and Traditional Literature Searching in Evidence Syntheses Using Four Case Studies. *Cochrane Evid Synth Methods*. 2025;3(6):e70050. 10.1002/cesm.70050.
5. Bernard N, Sagawa Y Jr, Bier N, et al. Using artificial intelligence for systematic review: the example of Elicit. *BMC Med Res Methodol*. 2025;25:75. 10.1186/s12874-025-02528-y.
6. Nye BE, Li JJ, Patel R et al. A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature. *Proc Conf Assoc Comput Linguist Meet*. 2018; 2018:197–207.

7. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. *Sci Rep*. 2024;14(1):6086. 10.1038/s41598-024-56706-x.
8. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag*. 2009;45(4):427–37. 10.1016/j.ipm.2009.03.002.
9. National Center for Biotechnology Information. Entrez Programming Utilities Help. NCBI Bookshelf. <https://www.ncbi.nlm.nih.gov/books/NBK25501/>
10. Fonferko-Shadrach B, Strafford H, Jones C, et al. Annotation of epilepsy clinic letters for natural language processing. *J Biomed Semant*. 2024;15(1):17. 10.1186/s13326-024-00316-z.
11. Bada M, Vasilevsky N, Baumgartner WA, Haendel M, Hunter LE. Gold-standard ontology-based anatomical annotation in the CRAFT Corpus. *Database (Oxford)*. 2017;2017:bax087. 10.1093/database/bax087.
12. Brooke J. SUS: a ‘quick and dirty’ usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. *Usability Evaluation in Industry*. London: Taylor & Francis; 1996. pp. 189–94.
13. Gronier G, Baudet A. Psychometric evaluation of the F-SUS: creation and validation of the French version of the System Usability Scale. *Int J Hum Comput Interact*. 2021;37(16):1571–82. 10.1080/10447318.2021.1898828.
14. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. *J Usability Stud*. 2009;4(3):114–23.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementarycontent.docx](#)