

Tracing and Filtering Unsupported Biomedical Extractions in a 55.5-Million-Relation Corpus

Input recovery, model replay, source composition and selective lexical safeguards

Narendra Kumar | Fliar BI | Preprint v2 | 15 September 2026

This preprint reports computational experiments and has not been peer reviewed. No independent clinical annotation is claimed.

Abstract

Background. Automated biomedical relations can reproduce source-selection and generation errors at scale. **Methods.** We audited 55,545,265 HCC v1.0 rows, standardized longitudinal labels by source, and extended an initial temporal diagnostic with 1,000 new sampled documents. Archived extractor inputs were recovered for 1,012 of 1,030 documents including discovery cases. We replayed a retained checkpoint on 70 new inputs and evaluated a fixed lexical rejection rule. **Results.** The unspecified-study share rose 22.13 percentage points in 2020, versus 1.92 with fixed source weights. Among 299 recovered pre-2019 holdout inputs, no tested COVID-family term occurred, yet the released outputs contained 380 explicit COVID-family rows. The gate rejected all 380. Replay generated 51 such rows from 30 held-out temporal inputs and matched all released tuples for 62 of 70 replayed documents overall. General-stratum row retention ranged from 77.2% to 86.3%. One of 829 deliberately mismatched-input controls passed. In the post-2020 comparison, 56 of 529 COVID-labelled rows also lacked the tested term family in the archived input. **Conclusions.** Input-level replay localizes an unsupported-generation failure in inspected cases; a transparent lexical safeguard suppresses these outputs with measurable loss of coverage. Lexical retention is not semantic accuracy or clinical validation.

Research contribution

The contribution is a measured failure analysis and a tested abstention mechanism: released counts, archived-input hashes, retained-model replay and holdout coverage are linked in one audit. We do not estimate treatment effects, disease incidence or the historical emergence of an infection.

Keywords: biomedical text mining; dataset audit; source composition; temporal inconsistency; publication metadata; research reproducibility.

1. Introduction and research questions

A disease-factor extraction record describes what a processing system emitted from a document. It does not, by itself, establish that the relation is present in the original text, correctly interpreted, or supported by a valid clinical study. A sufficiently large corpus can therefore yield precise-looking but substantively invalid summaries. The key measurement question is what each variable actually represents.

Biomedical semantic repositories such as SemMedDB demonstrate the utility of structured relations extracted from literature [2]. Data-statement approaches emphasize documenting the populations and processing choices behind NLP resources [3]. The present study applies these concerns to a mixed literature-and-web release. We evaluate HCC as delivered to researchers, rather than evaluating a newly trained model or treating model output as a clinical evidence base.

We ask four initial questions: How does source composition influence longitudinal label patterns? How do assigned study categories align with stored publication tags? Do time-sensitive disease names reveal inconsistencies that file checks cannot detect? How do lexical sparsity, repeated records and whole-period context fields affect reuse? Analyses are exploratory. A written protocol preceded this new full-release pass, but earlier HCC summaries existed; this was not a preregistered study.

The revision additionally asks whether original extractor inputs can localize the temporal anomalies and whether a fixed rejection rule reduces unsupported output on new cases. Faithfulness to input and factual truth are distinct constructs in generative NLP [10]. Abstention is an established response to uncertain predictions [11]; our gate is a deterministic lexical rule, not a trained SelectiveNet model or a demonstrated improvement in clinically adjudicated risk. We do not claim that hallucination detection or abstention itself is novel.

1.1 The released data

Component	Rows	Variables	Observation unit
Extraction records	55,545,265	30	One automated relation record
Documents	25,981,473	19	One transformed document identifier
Disease-factor pairs	40,819,026	13	One literal string pair
Disease terms	6,576,017	6	One literal disease-field value
Factor terms	18,848,567	6	One literal factor-field value

The release contains 22 compressed data files and six documentation/sample files, totaling 4,164,360,167 bytes [1,9]. The primary temporal field ranges from 1 January 2010 to 13 May 2026. Unknown dates affect 484,813 extraction records and 238,977 documents. Disease and factor strings are not normalized clinical concepts. Counts of distinct strings must not be presented as counts of recognized diseases or exposures.

The current audit verified the complete extraction, document, pair and entity counts against the released files. Source records with a public PubMed identifier can be linked externally; original source URLs, full input passages and sentence offsets are not supplied in the release.

2. Methods

2.1 Data generation and scope of reproducibility

The underlying workflow collected and downloaded biomedical literature records and web documents, selected medical text, and processed text into disease-factor relations. Available implementation files describe a two-stage FLAN-T5 classifier/extractor trained from model-generated examples, followed by a separate FLAN-T5 study/evidence classifier [4]. The label-generation scripts identify a Gemma model as the teacher. These details describe the recovered implementation, not an independently reconstructed history of every exported row.

The recovered extraction service truncates input to at most 768 tokens and uses greedy generation. Study/evidence classification uses at most the first 900 characters and 320 tokens of its supplied text. Thus, full-text methodological details can be absent from the classifier input. The teacher prompt itself couples evidence categories to study design. Historical student-versus-teacher scores are not used here as clinical accuracy estimates. The exact historical model revisions, all collection exclusions and lineage from every input passage to every released row remain incompletely documented.

2.2 Release-level checks and units

We read the fixed compressed release files without skipping malformed rows, verified their checksums, and checked unique identifiers, dates, document joins and declared extraction totals. The 2010-2025 main window includes 53,182,454 records. Partial 2026 and undated records are reported separately. Each linked document has one assigned study category repeated across its extraction rows; publication-tag comparisons therefore use documents, not extraction rows.

2.3 Temporal summaries and source standardization

For year y and source type s , let $p(y,s)$ be the fraction of extraction records carrying a specified label. Fixed weights $w(s)$ are the pooled 2010-2019 extraction shares among source types represented in every year of 2010-2025. The standardized rate is:

$$\text{standardized rate}(y) = \text{sum over } s \text{ of } w(s) * p(y,s)$$

Five source types satisfy common support: literature index, news media, other, press release and trade press. Promotional and unresolved categories are excluded from this standardization, removing 689,964 main-window rows. We also report common-source crude rates, so the effects of restricting the analysis set are visible. Fixed weights do not hold individual publishers, topics or model accuracy constant.

A symmetric Kitagawa decomposition separates changes in common-source crude rates into source-share and within-source-rate components [6]. COVID-term and released term-flag exclusions are sensitivity analyses, not validated cleaning rules. A second sensitivity panel retains 3,637 journals with at least 20 corpus documents in every year from 2010 through 2025, using fixed 2010-2019 document weights.

2. Methods (continued)

2.4 Publication-tag concordance and source checks

Eligibility for publication-tag comparisons requires a positive PubMed identifier and a populated stored publication-type list: 15,368,143 documents. The reference is the presence of the literal Randomized Controlled Trial tag. We compare it with any and all extraction records assigned the randomized-trial category within a document; these coincide because every document has one tier in this release. Neither the reference nor an absent tag is an adjudicated clinical study-design assessment. PubMed types are bibliographic fields and may change [5].

We selected 30 documents from each of four tag-by-label strata by ranking MD5(document identifier plus a fixed audit seed). This deterministic, balanced audit sample deliberately overrepresents uncommon cells. Current PubMed XML was retrieved for all 120 identifiers. We compared the stored RCT flag and complete publication-type sets with the retrieved fields. Results are counts for the audited records; they are not an unweighted population estimate of extraction accuracy.

2.5 Temporal lexical control

An additional exploratory diagnostic identified disease-field strings explicitly matching COVID-19 or SARS-CoV-2, including simple hyphen/space variants. We used dates before 2019 as a conservative temporal flag: WHO announced the name COVID-19 in February 2020 [8]. This test flags incompatible recorded-date/label combinations; it does not diagnose whether the origin is model output, source-content updates, date assignment or incorrect linkage.

From the flagged pre-2019 documents with a PMID, we selected 30 different documents using a second fixed hash seed and one extraction row per document. Current publication-year fields, titles and abstracts were retrieved. We tested the same literal terms in title plus abstract. Absence from those fields is not absence from every version of a full article or from the actual text originally supplied to the model. This diagnostic was added after observing pre-pandemic matches in the broader COVID sensitivity filter; it is explicitly exploratory.

2.6 Sparsity, redundancy and uncertainty

We tabulated literal pair support and the conditional distribution of evidence categories given study categories. Entropy reduction summarizes dependence between two assigned fields; it does not quantify clinical information. We counted early-appearing pairs with later records to identify the use of future information in whole-period context totals [7]. Within-document duplicate extraction signatures were inspected in a deterministic 1/256 hash-selected document sample.

All primary rates describe the released corpus itself. We do not attach binomial confidence intervals to 55.5 million dependent extraction rows, and do not interpret large-sample significance tests as protection against selection or label error. The observed decomposition is algebraic and descriptive, not causal. Detailed predicates, denominators, reference checks and reviewer worksheets accompany the technical supplement.

2.7 Archived-input recovery and held-out correction test

After the initial 30-document temporal audit, we froze a new deterministic document sample before inspecting its inputs: 300 PMID-linked explicit-COVID documents dated 2010-2018, 300 such documents dated 2020-2025, and 100 documents from each of four PMID-status-by-period strata. The general periods were 2010-2018 and 2019-2025. Discovery cases and already selected documents were excluded from later sampling frames. Selection ranked document identifiers with fixed MD5 seeds; exact frame counts and seeds accompany the supplement. This was a new evaluation set, not a preregistered confirmatory trial.

We joined released document hashes to retained provenance and recovered archived text prepared for extraction. All 1,030 identifiers resolved; 1,012 had an archived input. The 1,044 matching archive records contained repeat copies but no document with conflicting input text. We fingerprinted each unique input and kept full texts outside the public package. Missing cases remain visible in every denominator. These artifacts improve lineage for the sample but do not prove the historical checkpoint used for every released row.

Correction rule and comparison

The fixed candidate normalizes case and Unicode form, tokenizes ASCII letters and digits, and requires both disease and factor to appear as complete contiguous token sequences in the archived input. Rows failing either check are withheld in an experimental derivative and retained with a reason in an audit sidecar. Dates and entity labels are never rewritten. We compare this rule with the existing diagnostic: raw normalized substring presence or at least 60% term-token overlap. The recovered client recorded that diagnostic but did not reject failing rows.

We report retained-row coverage, not semantic precision. Each general stratum is reported separately; the balanced sample is not population-proportional. Percentile intervals use 2,000 document-cluster bootstrap draws within each recovered-input stratum. They are conditional on recovery and omit uncertainty from missing sources, lexical definitions and corpus selection. Deliberately mismatched-input controls cyclically permute documents within each general stratum; chance entity overlap remains possible, so their pass rate is not clinical specificity.

Checkpoint replay

Three discovery inputs were replayed first. A separate seed then selected 30 temporal-holdout, 20 COVID-comparison and five documents from each general stratum for a 70-document replay. The retained checkpoint used float16, greedy decoding, at most 768 input tokens and 256 output tokens, one document per call. Parsed tuples were compared after the release normalization. No parameters were tuned on the new replay outputs. Current checkpoint fingerprints and the exact replay sample are supplied; historical hardware and row-specific model hashes remain unavailable.

3. Results: coverage and changing source composition

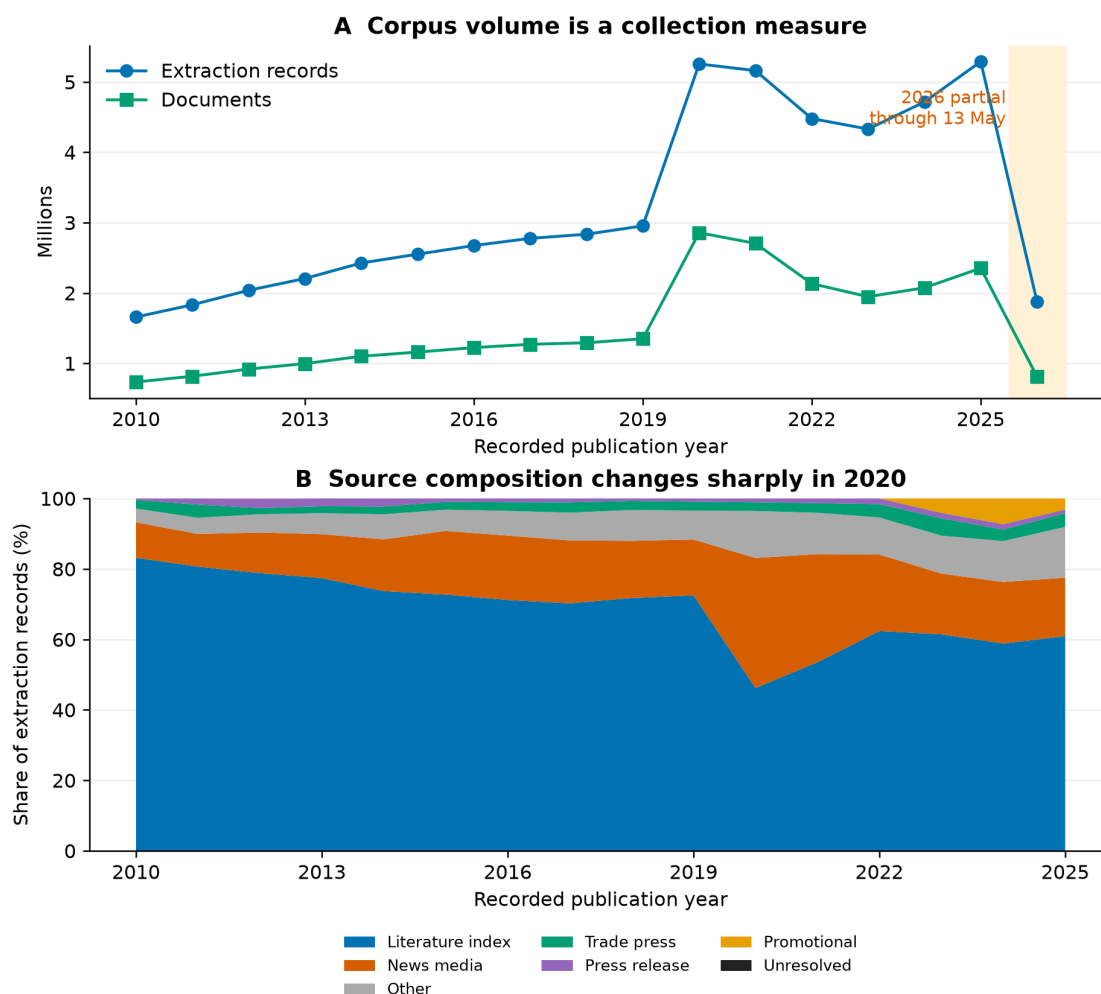


Figure 1. A: Complete-release annual counts; the shaded 2026 endpoint is partial through 13 May. B: Source-type shares for 2010-2025. Unknown dates are excluded from both plots. Counts describe retained corpus records, not all publications or health events. Source data: `claims_year_source.csv` and `document_year_source.csv`.

Extraction volume rose from 2,953,981 records in 2019 to 5,255,612 in 2020, a 77.92% increase. News-media records increased from 467,988 to 1,938,742. Their share rose from 15.84% to 36.89%, while literature-index share fell from 72.49% to 46.21%. Consequently, the mix of material contributing to an annual percentage changed substantially.

Across the complete release, 35,390,352 extraction rows and 15,368,148 documents carry a PMID. Those document identifiers map to the same number of distinct PMIDs. This supports clear bibliographic linkage for that subset; it does not establish representative coverage of PubMed, all biomedical literature or the web.

3.1 The apparent 2020 study-label shock is composition-sensitive

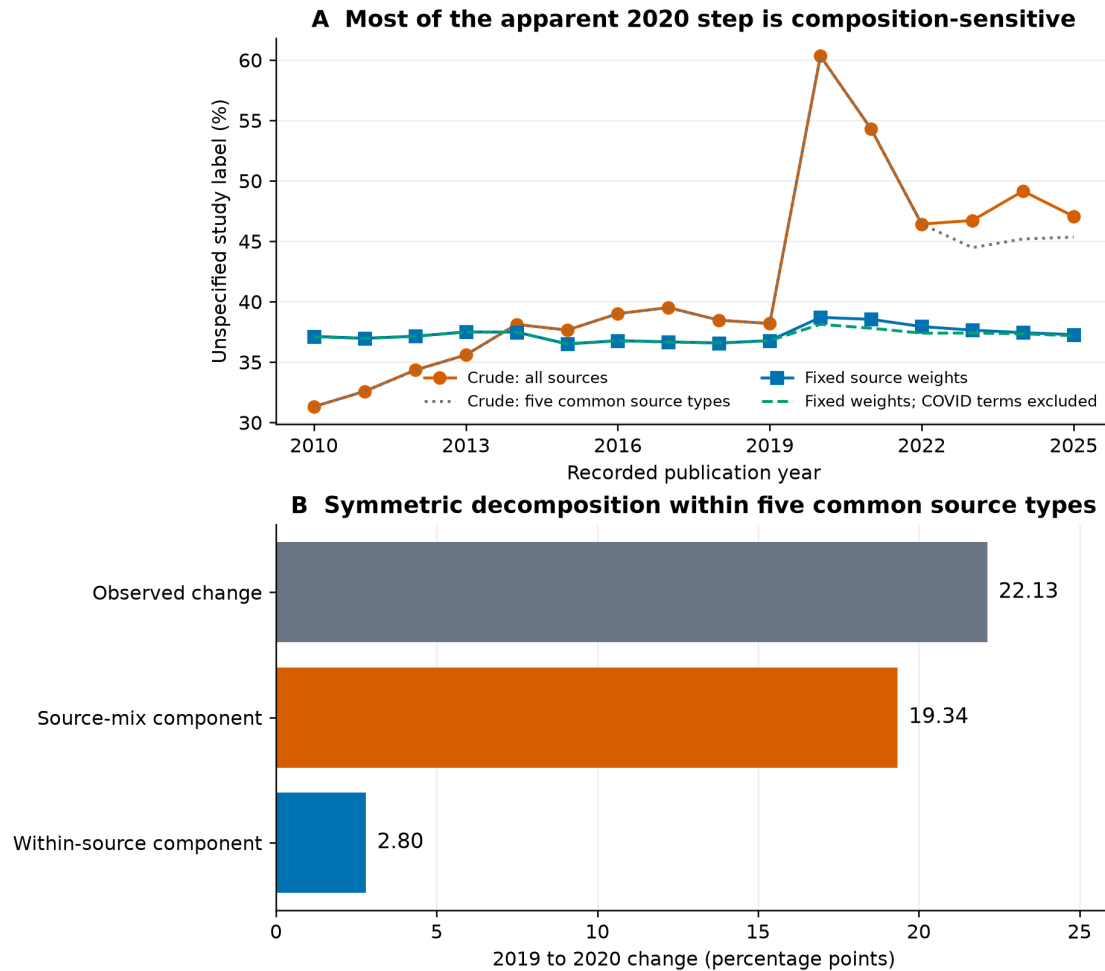


Figure 2. A: Unspecified-study label shares under crude, common-source and fixed-weight summaries. B: Symmetric decomposition of the 2019-2020 common-source change. Five source types contribute; fixed weights use 2010-2019. Components sum to the observed change. Neither component is a causal effect. Source data: `annual_sensitivity.csv`, `fixed_source_weights.csv` and `decomposition.csv`.

The all-source unspecified share increased by 22.13 percentage points, from 38.19% to 60.33%. Its fixed-source-weight counterpart increased by 1.92 points, from 36.78% to 38.70%. In the symmetric common-source decomposition, 19.34 points lie in the source-mix component and 2.80 points in the within-source component; the former is 87.4% of the observed increase. These components differ from the fixed-baseline contrast because their weighting formulas differ.

After excluding the broader COVID lexical family, the crude 2020 unspecified share remains 51.22%, whereas its standardized value is 38.13%. Removing COVID matches alone therefore does not remove the composition problem. None of these changes measures the quality of real-world research over time.

3.2 Assigned randomized-trial labels are not interchangeable with tags

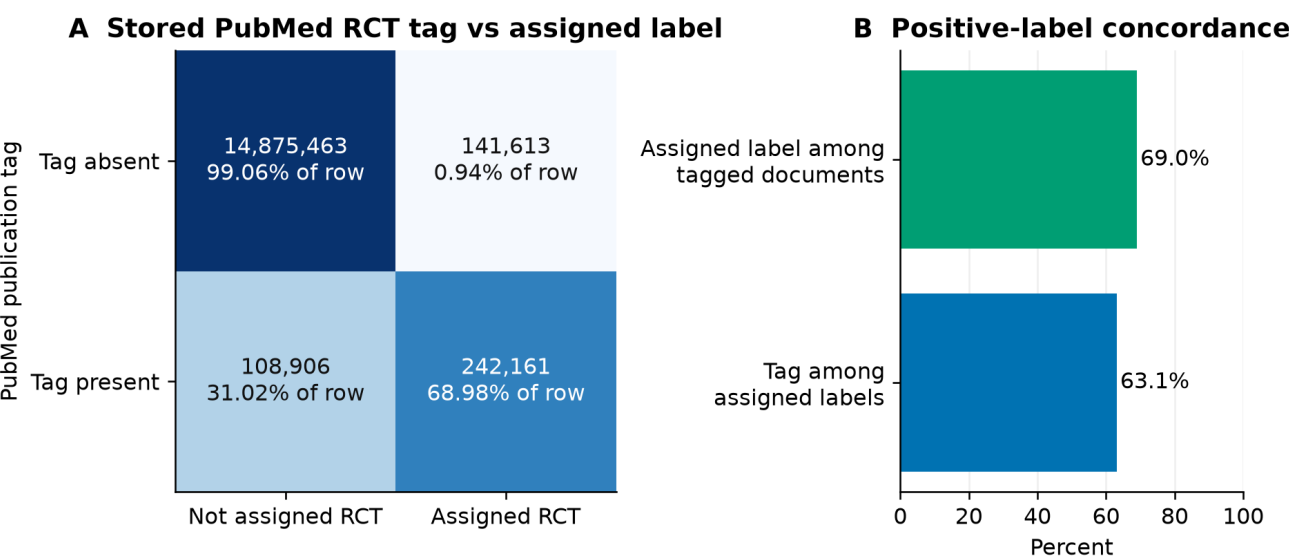


Figure 3. Document-level comparison among 15,368,143 PMID-linked records with nonempty stored publication types. Cell percentages use the PubMed-tag row denominator. The labels “tag absent” and “not assigned” do not mean a study was clinically adjudicated as non-randomized. Source data: rct_concordance.csv.

Of 383,774 documents assigned the randomized-trial category, 242,161 carry the stored RCT tag (63.10%); 141,613 do not. Among 351,067 tagged documents, 108,906 lack the assigned RCT category, leaving positive-tag coverage of 68.98%. The overall agreement is 98.37%, but always assigning “tag absent” would already agree in 97.72% of cases. Overall accuracy alone obscures the positive-class discrepancies.

Tagged documents not assigned the RCT category were distributed across unspecified (46,383), clinical (31,767), observational (14,588), preclinical (11,619) and review/meta-analysis (4,549) labels. These differences could arise from truncated inputs, label-definition differences, missing or evolving publication indexing, or model errors. The comparison does not isolate their contributions.

Point-in-time source verification

All 120 audit PMIDs were retrieved. The stored RCT flag matched the current flag in 119 cases; the complete type set matched in 116. One stored tag-absent, model-RCT case had acquired an RCT tag in the current metadata. That example shows why an apparent disagreement with a frozen tag must not automatically be called a model false positive. The audit is a metadata reproducibility check, not an independent expert assessment of the extracted disease-factor statements.

3.3 A temporal lexical control reveals incompatible date-label pairs

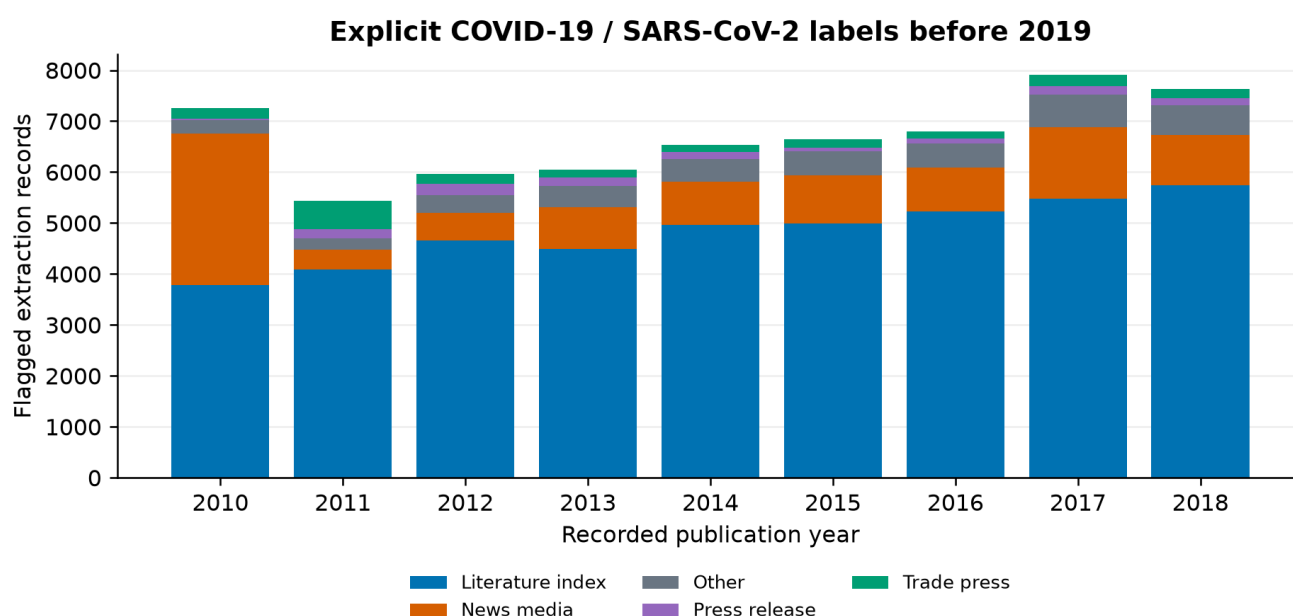


Figure 4. Explicit COVID-19 / SARS-CoV-2 disease labels attached to 2010-2018 document dates. The conservative predicate, source counts and exact sample are supplied in the supplement. Generic coronavirus mentions are not included in this diagnostic. Source data: strict_covid_year.csv.

The diagnostic flagged 60,235 records across 50,568 documents dated 2010-2018, or 0.287% of all extraction rows in those years. Of the flagged records, 43,334 have a PMID. The test identifies a concrete temporal incompatibility; it is not evidence that COVID-19 existed in those years.

In the targeted sample of 30 PMID-linked flagged documents, every current journal year agreed with the stored year, all 30 had an abstract, and none of the current title/abstract pairs contained the tested terms. Thus, in these cases, simply replacing the stored year with the current PubMed journal year would not resolve the anomaly. The archived-input and replay experiments below investigate this origin for a new sample; full revision history is still unavailable.

Two inspectable examples are [PMID 20472450](#) (2010; irradiation and chemical-material properties) and [PMID 23527560](#) (2013; energy-value prediction for broiler diets). Each received a COVID-19 disease label in HCC. These cases illustrate why a valid identifier and matching publication year do not establish semantic validity. We have not classified every flagged record as a confirmed hallucination.

3.4 Archived inputs localize an unsupported-generation failure

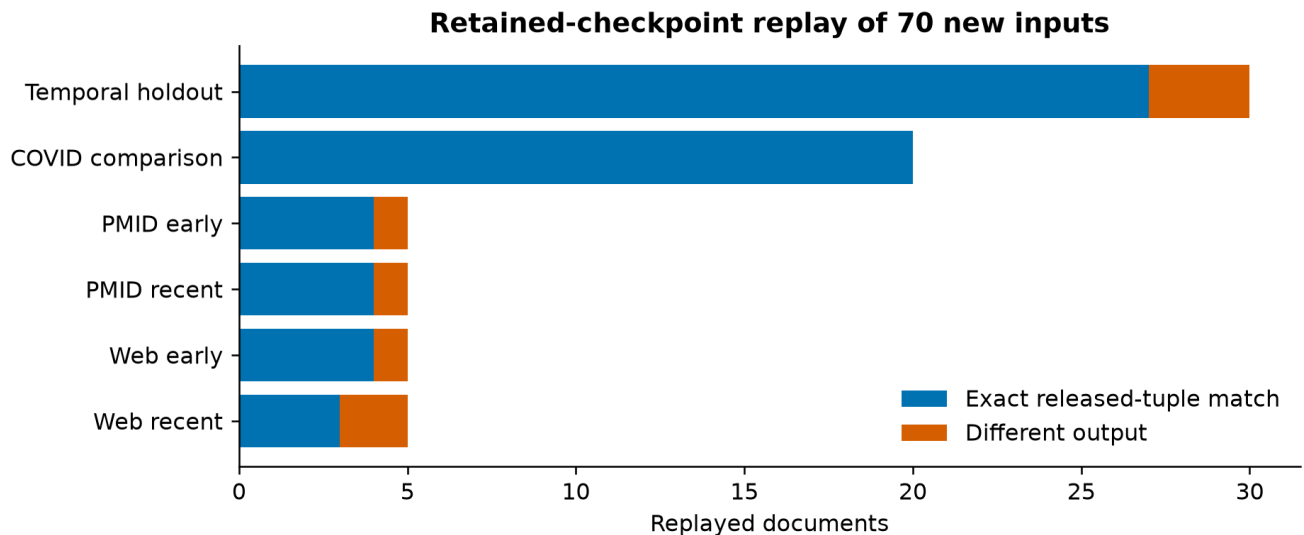


Figure 5. A retained-checkpoint replay of 70 newly sampled archived inputs reproduces all released five-field tuples for 62 documents. The eight discrepancies are reported, not discarded. Matching is reproducibility, not clinical accuracy. Source data: `replay_sample.csv`, `replay_comparison.csv` and `replay_summary.csv`.

All 30 discovery inputs were recovered. None contained the tested COVID-family terms. The three initial replay examples reproduced the four released COVID-19 relations exactly, including the material-irradiation and broiler-diet cases described above. The checkpoint therefore generates those disease labels from input passages without the corresponding terms; replacing a publication year does not repair this observed behavior.

In the separate 70-document replay, 62 outputs matched the complete released tuple lists after normalization. The temporal stratum matched in 27 of 30 documents; the retained model emitted 51 COVID-family rows from those 30 inputs. No tested COVID-family term occurred in any of them. All 51 replayed rows fail the proposed disease-phrase requirement. This directly demonstrates unsupported label generation in this experiment, rather than only comparing two metadata fields.

The eight replay discrepancies limit claims about exact historical reproducibility. Different numerical execution, processing revisions, tuple truncation or retained checkpoint identity may contribute; these causes were not isolated. We did not demonstrate that every temporal anomaly has the same origin. The experiment does not establish why training produced this behavior or validate unrelated outputs.

3.5 A lexical safeguard suppresses anomalies with loss of coverage

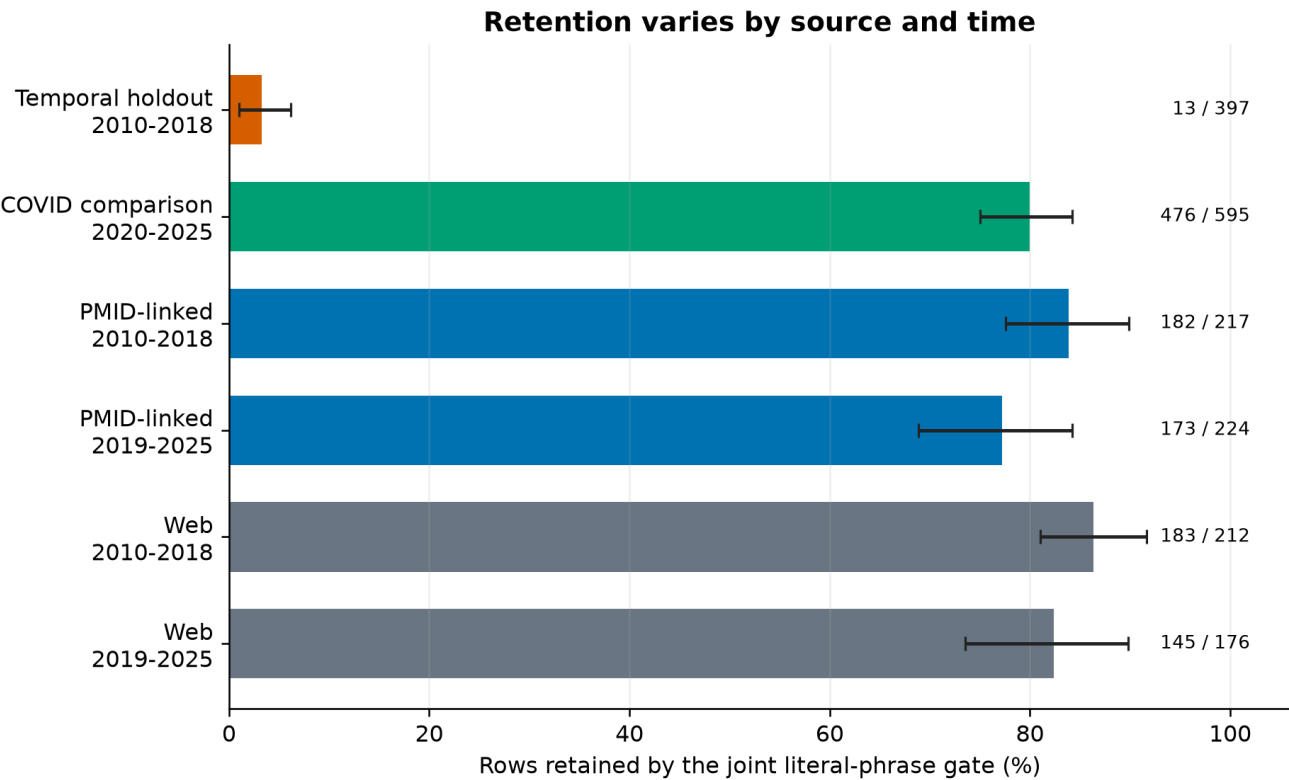


Figure 6. Joint disease-and-factor literal-phrase retention among recovered unambiguous inputs. Bars use extraction rows; 95% percentile intervals resample whole documents within each stratum (2,000 draws). Numerators and denominators appear at right. Missing inputs are excluded from these conditional rates and reported in the supplement. Source data: `grounding_summary.csv` and `grounding_retention_intervals.csv`.

Among 300 new pre-2019 PMID-linked temporal cases, 299 inputs were recovered. Their 397 released rows include 380 explicit COVID-family labels; none of the 299 inputs contains the tested term family. The gate rejects all 380 while retaining 13 other rows. This validates rejection of this observed failure class; it does not establish the correctness of the 13 retained relations.

General-stratum retention is 182/217 (83.9%) for early PMID-linked records, 173/224 (77.2%) for recent PMID-linked records, 183/212 (86.3%) for early web records and 145/176 (82.4%) for recent web records. The legacy token-overlap diagnostic retains more rows in each stratum. We cannot tell without adjudication how much additional rejection removes errors versus valid paraphrases or aliases. This is a measured coverage cost, not evidence of a corresponding gain in semantic precision.

3.6 Date filtering alone misses unsupported recent labels

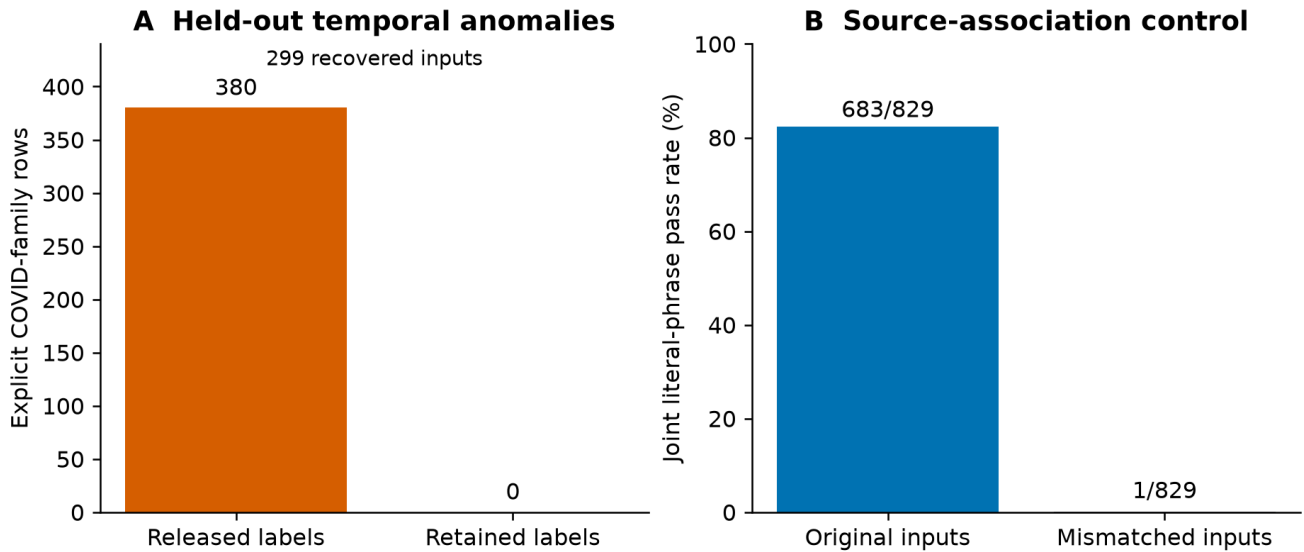


Figure 7. A: The 380 explicit COVID-family rows in recovered temporal holdout inputs are all withheld by the candidate gate. B: 683/829 rows pass on their own general-sample inputs versus 1/829 on deliberately mismatched inputs. The pooled general sample is an artificial balanced mixture. Neither panel estimates clinical precision, recall or specificity. Source data: `grounding_audit.csv` and `mismatched_input_controls.csv`.

In the post-2020 comparison, 293 of 300 inputs were recovered. Of 529 explicit COVID-family rows in that assessed set, 56 come from 41 inputs containing none of the tested family terms. The problem is therefore not confined to pre-pandemic dates. A rule that only removes pre-2019 COVID labels leaves these recent cases untouched. The joint literal-phrase rule retains 421/529 explicit-COVID rows and 476/595 rows overall in this comparison; this includes rejection beyond the 56 family-absent rows.

Across the four general strata, the new gate passes 683/829 original-input rows (82.4%) and one of 829 mismatched-input rows (0.12%). The legacy diagnostic passes 770/829 original-input rows (92.9%) and two mismatched-input rows. The controls show sensitivity to which text is attached to an output, but are easier than genuine extraction errors: unrelated documents rarely share both entities. They cannot replace human annotations.

Explicit tests also demonstrate the safeguard's limits: entity phrases can occur in a negated claim or in unrelated sentences and still pass. Nine mechanical tests cover missing entities, empty strings, token boundaries, punctuation, alias abstention and these known semantic limitations. A passed gate means only that literal entity support is present. An independent relationship-level evaluation remains necessary before clinical reuse.

3.7 Assigned evidence categories provide limited independent information

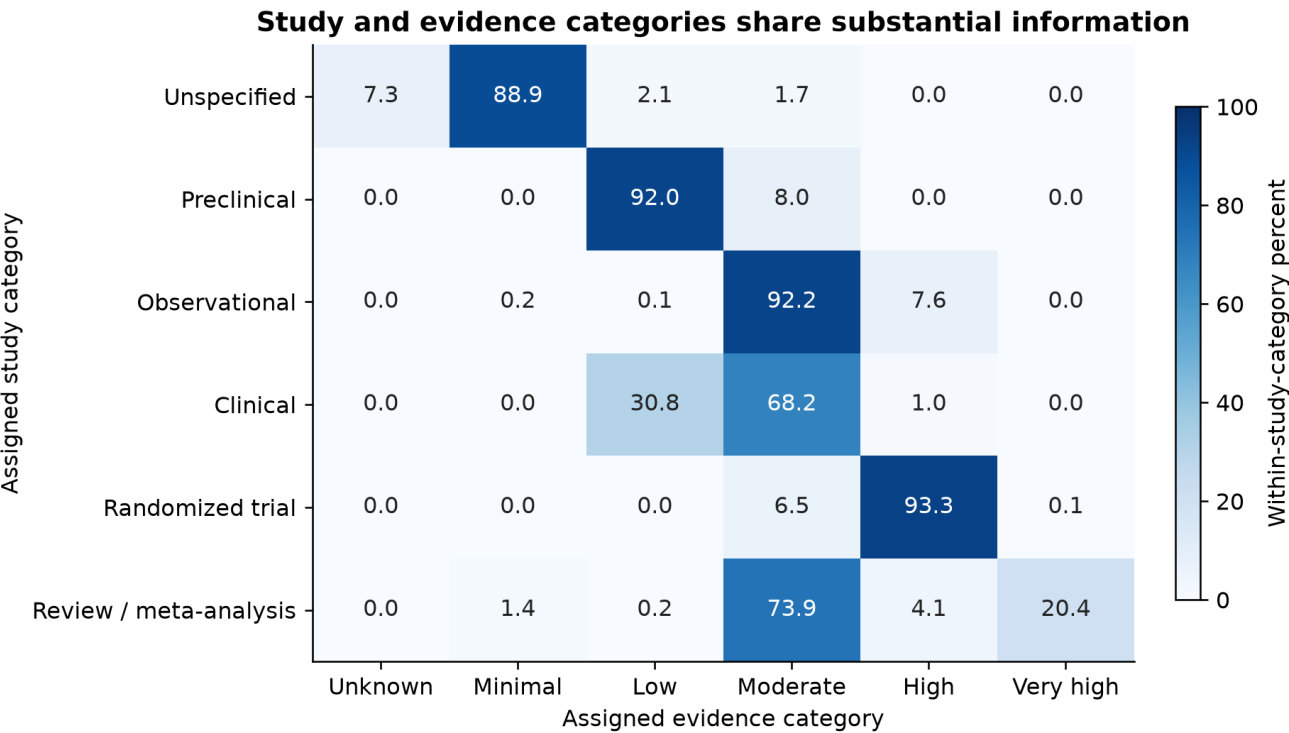


Figure 8. Evidence-category percentages conditional on assigned study category, 2010-2025. These are two outputs of the same model workflow. Unknown is retained. Percentages do not represent clinical certainty or an independently validated evidence scale. Source data: tier_strength.csv.

A lookup that chooses the most common evidence category for each study category reproduces 86.46% of evidence labels. Knowing the study category reduces the empirical entropy of the evidence category from 1.948 to 0.628 bits, a 67.76% reduction. Randomized-trial labels co-occur with “high” evidence in 93.3% of records, while preclinical labels co-occur with “low” in 92.0%.

This dependence is consistent with the recovered labeling prompt, which explicitly links evidence strength to design. It is not a demonstration that either field is clinically correct. Treating a study-category filter and an evidence-category filter as two independent corroborating signals would double-count much of the same assigned information.

The numeric scores supplied with the dataset encode category names. Their ordering and distances have not been validated as measurement properties. Averaging those numbers over time should not be interpreted as an average quality of scientific evidence. Where a research question needs a clinical evidence assessment, article-level evaluation and an explicit, independently applied rubric are required.

3.8 Literal pair sparsity and temporal context constrain reuse

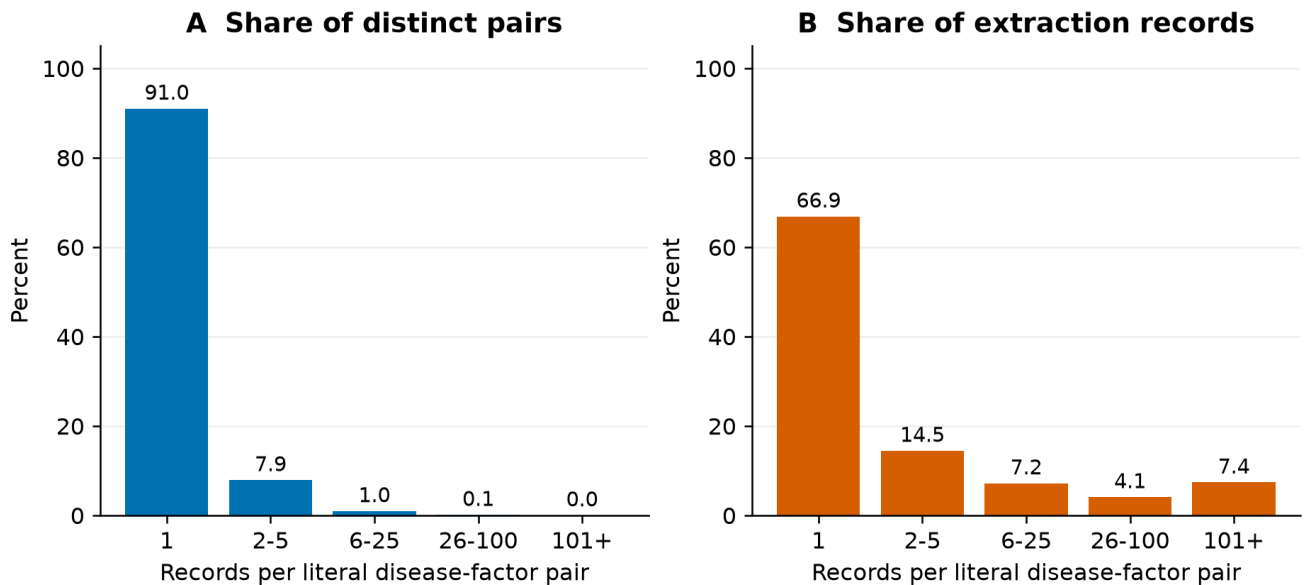


Figure 9. Frequency distribution of literal disease-factor string pairs. A singleton is one extracted pair observed once in the entire release, not necessarily a unique biomedical concept. Raw aliases are not merged. Source data: pair_support.csv.

Of 40,819,026 literal pairs, 37,143,799 occur once (91.00%). These singleton pairs account for 66.87% of extraction rows. The release contains 6,576,017 distinct disease-field strings and 18,848,567 factor strings, so lexical variation and extraction noise can fragment clinically related concepts. No synonym-normalized network is evaluated here.

Only 348 literal pairs satisfy the released threshold of at least 25 increasing-direction and 25 decreasing-direction records. This threshold detects opposite assigned directions; it does not establish conflicting scientific findings. Different outcomes, populations, formulations, time horizons and extraction errors can produce opposite labels.

Context fields and future information

Among 19,697,186 pairs first appearing in 2010-2019, 1,111,107 (5.64%) also have later records. Their supplied whole-release totals therefore include information unavailable at a 2019 cutoff. The same concern applies to disease and factor totals accumulated across the release. For prospective prediction, rebuild such features within the training window, after document-level splitting. This study demonstrates an exposed future-information pathway, not a measured amount of predictive-performance inflation [7].

3.9 Internal integrity is strong, but it is a different property from validity

Check	Observed result	What it establishes
Extraction identifiers	55,545,265 unique / 55,545,265 rows	No identifier duplicates
Document identifiers	25,981,473 unique / 25,981,473 rows	No identifier duplicates
Extraction-to-document join	0 unmatched identifiers	All exported references join
Per-document extraction totals	0 mismatches	Declared totals reconcile
Source labels and document years	0 within-document inconsistencies	Exported grouping is internally coherent
Stored year vs PubMed year	99.6328% agreement among 15,368,148	Agreement of two recorded year fields

All supplied per-document row totals equal the observed counts, and source types agree between extraction and document records. Every document has one study category across its rows. These checks justify the computational joins and document weighting used here. They do not make the rows statistically independent or validate medical content.

Stored document year agrees with PubMed publication year in 15,311,721 of 15,368,148 comparable records. There are 2,436 differences exceeding one year. PubMed can expose electronic and print dates [5], so a year difference need not be an error. We did not adjudicate all date discrepancies. Conversely, the temporal lexical audit shows that year agreement can coexist with an implausible disease label.

In a deterministic hash-selected sample of 101,004 documents and 215,756 extraction records, 60 rows repeat a within-document signature of disease, factor, relation, direction, stance, study category and evidence category (0.0278%). This narrow diagnostic does not assess syndicated pages, near duplicates, repeated underlying studies or semantic duplicates across different document identifiers.

What remains outside these checks

The public release does not expose inference passages, token offsets, collection-time snapshots or complete model-version lineage. This revision recovers archived inputs for 1,012 sampled documents; those private-source artifacts do not resolve provenance for the full corpus. A distinct document identifier can still represent repeated content. Source labels are coarse and partly heuristic. Consequently, the low identifier/signature duplication counts cannot establish that every contributing document is an independent scientific observation.

3.10 A fixed journal panel gives a narrower longitudinal view

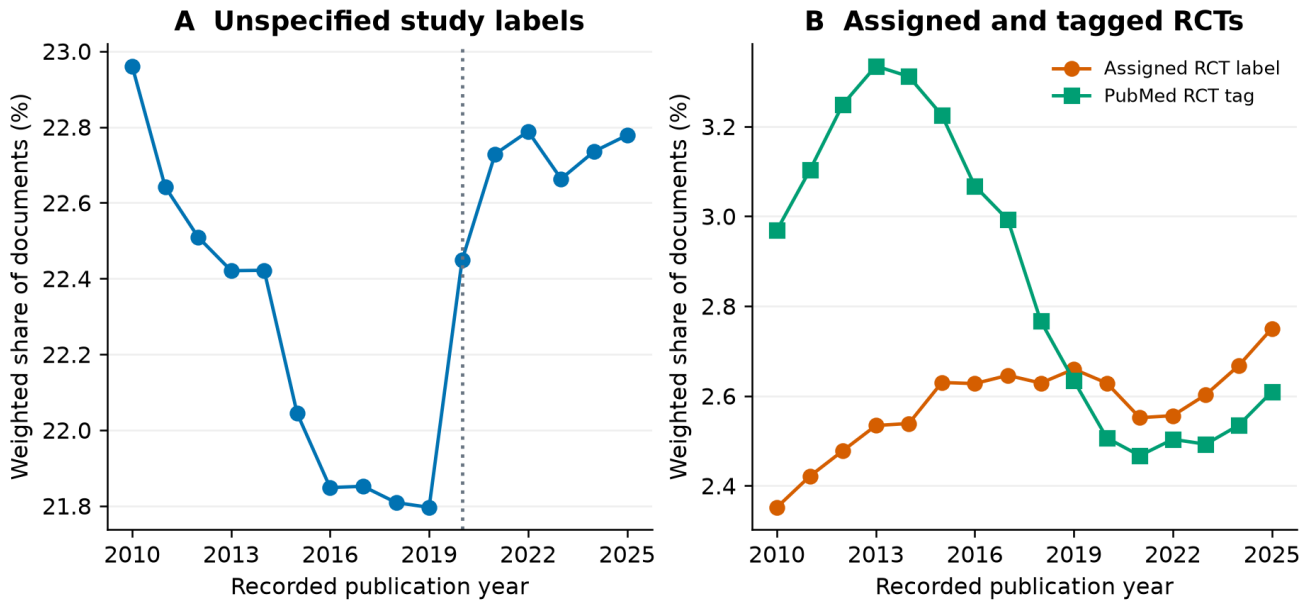


Figure 10. A balanced set of 3,637 journals with at least 20 retained corpus documents in each year 2010-2025; weights are fixed to pooled 2010-2019 document counts. This panel contains 10,626,480 of 14,907,557 eligible PMID-linked journal documents in that window. It is a selected panel, not representative sampling. Source data: journal_annual.csv and balanced_journal_trends.csv.

Within this panel, the unspecified-label share moves from 21.80% in 2019 to 22.45% in 2020, an increase of 0.65 percentage points. This is substantially smaller than the 22.13-point all-source extraction-row increase. The estimates have different units, coverage and weights; the panel is a sensitivity analysis rather than a replacement population estimate.

The document-weighted causal-label share rises from 3.89% to 4.11% within the panel, whereas the all-source extraction-row share falls from 4.90% to 3.79%. The direction of a broad corpus trend is therefore not portable across these analysis choices. We do not infer that causal evidence strengthened or weakened in either population: these are model-assigned relation labels.

Requiring journals to appear in every year uses knowledge of the completed observation window and selects persistent journals. It can exclude emerging, discontinued, renamed or sparsely captured outlets. It also leaves within-journal topic mix, indexing practices and label accuracy uncontrolled. This panel should not be used as a future-prediction validation split.

4. Discussion

The corpus audit and input-level experiments establish complementary results. Source mix can dominate annual label changes, and assigned study categories disagree with bibliographic tags. Archived inputs and checkpoint replay additionally demonstrate unsupported disease-label generation in inspected temporal cases. A fixed lexical safeguard rejects those outputs on a new sample but also removes 13.7-22.8% of general-stratum rows. These are measured corpus and extraction properties, not disease mechanisms.

The combination matters. A clean export can support precise reproducible calculations while the measured constructs remain weak. The strong identifier, join and count checks enable auditing; they cannot substitute for semantic evidence. Likewise, adding bibliographic enrichment makes discrepancies inspectable but does not transform an automatically emitted disease-factor relation into a verified finding.

4.1 Supported uses

HCC can be used to study extraction behavior, design normalization experiments, evaluate source-sensitive summaries, and construct candidate cases for independent annotation. The released corpus can also support hypothesis generation when candidate links are subsequently checked against source text and appropriate scientific evidence. The results reported here justify these audit-oriented uses; they do not establish performance on clinical decision tasks.

4.2 Limits and follow-up validation

No independent expert annotation study was completed for this manuscript. The source checks verify metadata and limited lexical properties; they do not measure relationship precision or recall. The original review could not distinguish missing abstract context from unsupported generation. The recovered-input replay now establishes unsupported output in the replayed temporal cases, while missing inputs, current-versus-historical checkpoint differences and other cases remain unresolved. It does not identify which training or architectural choice caused the behavior. The audit sample strata are intentionally unequal in their population representation, and the temporal sample is selected specifically from flagged records.

A next semantic validation phase should expand exact-input recovery, select a fresh document-level probability sample across sources and years, obtain two independent annotations under a written rubric, and adjudicate disagreements. It should evaluate disease/factor validity, textual support, relation direction and study design separately. Test/train separation must be at the document level, with duplicate-aware grouping and time-local feature construction. An explicit comparison with an established extraction benchmark is still needed before making comparative performance claims.

5. Conclusion

HCC is computationally coherent but clinically unvalidated. This study moves from an observed temporal anomaly to archived-input recovery, retained-model reproduction and a measured rejection mechanism. The lexical safeguard removes the demonstrated unsupported labels on new cases, at a material coverage cost, and cannot establish relation entailment. Source-aware analysis and explicit abstention improve the transparency of reuse; independent semantic validation remains necessary.

Data availability, reproducibility and declarations

The audited HCC v1.0 files are publicly archived at [doi:10.5281/zenodo.22754767](https://doi.org/10.5281/zenodo.22754767), with an identical Figshare mirror at [doi:10.6084/m9.figshare.33752332.v1](https://doi.org/10.6084/m9.figshare.33752332.v1) [1,9]. The dataset was produced by [Fliar BI](#). A small example CSV is available from the project at [the Fliar BI HCC sample page](#). That existing sample is illustrative and must not be used to estimate corpus proportions.

The [Fliar BI study materials](#) include analysis scripts, compressed-input checksums, tabulated figure data, sampled correction outputs, source-audit identifiers, figures and a technical supplement. It reproduces analyses of the released data; the revision also supplies input fingerprints, replay outputs and the lexical gate. Full archived input texts and retained model weights are not redistributed, so independent regeneration of the input-level and model replay experiments requires access to those artifacts. Tabulated results and the gate are public; this is not end-to-end reproduction of original collection and training. Fresh PubMed checks may differ from the archived 15 September 2026 retrievals because the reference metadata evolve.

Authorship and assistance. Narendra Kumar is the designated author. Computational analysis and manuscript drafting used AI assistance under the project owner's direction. Fliar BI produced and distributes the audited dataset and conducted this evaluation; this producer-led audit is not independent external validation. No independent human expert annotation is claimed.

Ethics and licensing. The analyzed release is distributed under CC BY 4.0. We do not redistribute source abstracts in this review package. Public bibliographic identifiers allow linkage and must not be described as a guarantee of anonymization. The findings are not patient-level clinical research or medical advice. Repository deposit is not peer review.

References

- [1] Fliar BI. Health Claims Corpus (HCC), version 1.0. Zenodo (2026). Dataset. <https://doi.org/10.5281/zenodo.22754767>
- [2] Kilicoglu H, Shin D, Fiszman M, Rosembat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics* 28, 3158-3160 (2012). <https://doi.org/10.1093/bioinformatics/bts591>
- [3] Bender EM, Friedman B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6, 587-604 (2018). https://doi.org/10.1162/tac1_a_00041
- [4] Chung HW and colleagues. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research* 25, 1-53 (2024). <https://www.jmlr.org/papers/v25/23-0870.html>
- [5] National Library of Medicine. PubMed User Guide. Publication types and publication-date fields. Accessed 15 September 2026. <https://pubmed.ncbi.nlm.nih.gov/help/>
- [6] Kitagawa EM. Components of a Difference Between Two Rates. *Journal of the American Statistical Association* 50, 1168-1194 (1955). <https://doi.org/10.1080/01621459.1955.10501299>
- [7] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* (2023). <https://www.sciencedirect.com/science/article/pii/S2666389923001599>
- [8] World Health Organization. Naming the coronavirus disease (COVID-19) and the virus that causes it. Naming announcement: 11 February 2020. [https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-\(covid-2019\)-and-the-virus-that-causes-it](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it)
- [9] Fliar BI. HCC version 1.0. Figshare (2026). Identical release mirror. <https://doi.org/10.6084/m9.figshare.33752332.v1>
- [10] Maynez J, Narayan S, Bohnet B, McDonald R. On Faithfulness and Factuality in Abstractive Summarization. *ACL* (2020), 1906-1919. <https://aclanthology.org/2020.acl-main.173/>
- [11] Geifman Y, El-Yaniv R. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. *ICML, PMLR* 97 (2019), 2151-2159. <https://proceedings.mlr.press/v97/geifman19a.html>