

HCC input-level audit: technical supplement

Methods, denominators, source checks and independent-review protocol

Fliar BI | Preprint v2 | 15 September 2026

S1. Exact analysis sets and integrity checks

Analysis	Denominator / scope
Full extraction release	55,545,265 rows; 55,545,265 unique extraction identifiers
Document release	25,981,473 unique identifiers; sum of declared extraction counts equals 55,545,265
Complete-year main window	53,182,454 extraction rows, 2010-2025
Partial 2026	1,877,998 extraction rows, recorded dates through 13 May
Undated	484,813 extraction rows; 238,977 document records
Stored PMID links	15,368,148 documents / distinct PMIDs
Publication-tag comparison	15,368,143 documents with PMID and nonempty publication-type list
Common-source standardization	52,492,490 extraction rows over 2010-2025; 689,964 omitted
Balanced journal panel	3,637 journals; 10,626,480 documents
Within-document signature audit	101,004 documents; 215,756 rows; 60 repeated rows

All five released components were counted from their files. Summing the claim totals in the pair, disease and factor summaries separately yields 55,545,265 in each case. No release file was changed for these analyses. Null dates are retained as unknown; no inferred dates are introduced. The source type, claim count and recorded year agree between every linked document and its extraction records.

Exact checks and query outputs are supplied in `claim_integrity.csv`, `document_integrity.csv`, `join_integrity.csv`, `dimension_integrity.json` and `ingest_manifest.json`. “Verified” in those files denotes the specified computational check only. It never denotes clinical verification.

S2. Complete annual label summaries

Year	Records	Unspecified %	Fixed-source %	Causal-label %	High-category %
2010	1,658,774	31.32	37.13	4.24	4.08
2011	1,832,337	32.58	36.97	4.32	4.23
2012	2,038,310	34.35	37.14	4.41	4.21
2013	2,205,775	35.60	37.50	4.57	4.32
2014	2,426,457	38.12	37.48	4.67	4.34
2015	2,551,347	37.65	36.51	4.68	4.68
2016	2,674,136	39.01	36.77	4.83	4.61
2017	2,776,433	39.51	36.68	4.87	4.74
2018	2,835,369	38.47	36.58	4.98	4.84
2019	2,953,981	38.19	36.78	4.90	4.94
2020	5,255,612	60.33	38.70	3.79	3.26
2021	5,160,696	54.30	38.54	3.73	3.97
2022	4,479,274	46.42	37.94	4.57	4.86
2023	4,330,139	46.72	37.65	4.90	5.01
2024	4,714,766	49.16	37.44	4.79	4.99
2025	5,289,048	47.07	37.28	4.58	5.63

Percentages are extraction-record shares. “High-category” combines assigned high and very-high labels; it is not a count of independently assessed high-quality studies. Fixed-source rates use the same five source categories and baseline weights in each year. Partial 2026 and undated records are intentionally absent. All variants, common-source crude rates and denominators appear in `annual_sensitivity.csv`.

Fixed baseline weights

Source category	2010-2019 baseline weight
literature index	74.5017%
news media	14.9289%
other	6.7306%
press release	1.3640%
trade press	2.4747%

S3. Sensitivity analyses and decomposition

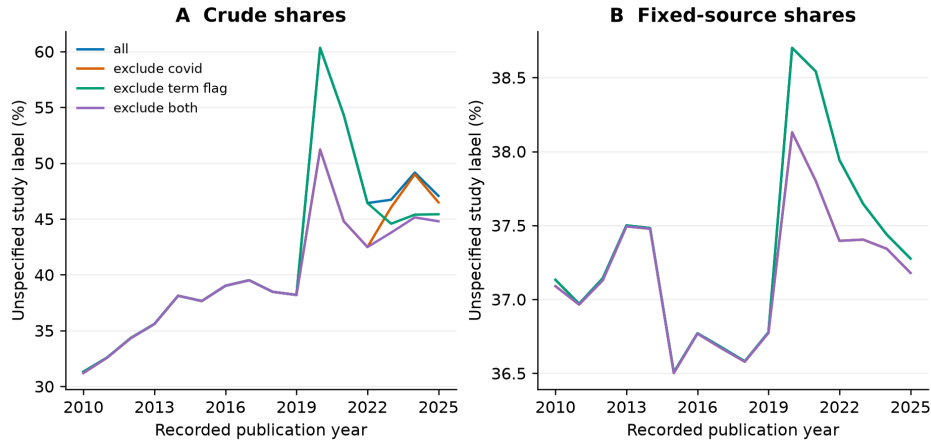


Figure S1. Term filters and source weighting answer different descriptive questions. The released lion/mane flag is not a validated classification of promotional content, and the COVID filter is a lexical rule rather than a controlled-vocabulary concept mapping.

The broad COVID sensitivity rule matches disease strings containing “covid”, or variants of SARS-CoV-2 / 2019-nCoV. It can miss synonyms and include unsuitable strings. The stricter temporal control in S5 is a separate predicate. The released term flag marks a factor-string pattern with “lion” preceding “mane”; its exclusion is a sensitivity check only.

Let $w_0(s)$, $w_1(s)$ be source shares at the two endpoints, and $p_0(s)$, $p_1(s)$ their label fractions. The symmetric decomposition is:

$$\begin{aligned} \text{Composition} &= \sum_s (w_1 - w_0) * (p_1 + p_0) / 2 \\ \text{Within-source} &= \sum_s (p_1 - p_0) * (w_1 + w_0) / 2 \\ \text{Total change} &= \text{Composition} + \text{Within-source} \end{aligned}$$

This is an identity for the observed weighted rates, not an estimate of a causal mechanism. It is applied only within the five common source types. All-source and common-source crude changes can differ, especially when promotional-source records emerge in later years.

Period	Label	Total pp	Mix pp	Within pp
2019-2020	unspecified	22.133	19.336	2.798
2019-2020	causal	-1.105	0.600	-1.705
2019-2020	high	-1.671	-1.244	-0.427
2022-2023	unspecified	-1.944	-1.571	-0.373
2022-2023	causal	0.523	-0.078	0.602
2022-2023	high	0.364	0.183	0.180
2010-2025	unspecified	14.029	13.945	0.084
2010-2025	causal	0.473	0.659	-0.186
2010-2025	high	1.729	-0.727	2.456

S4. Stratified publication-metadata check

Stored tag / assigned RCT	Frame documents	Audited	RCT flag matches	Type sets match
0 / 0	14,875,463	30	30	30
0 / 1	141,613	30	29	28
1 / 0	108,906	30	30	29
1 / 1	242,161	30	30	29

Here 1 means present/assigned and 0 means absent/not assigned. Each stratum contributes 30 documents regardless of its population size. Inclusion is determined by sorting MD5(doc_id + "HCC-audit-20260915"), with the document identifier breaking ties. Full identifiers, stored/current fields, retrieval timestamps and stratum sizes are in `pubmed_metadata_audit.csv`. No independent annotation labels were created by this check.

Observed changes in the retrieved reference

PMID 41052730. Stored: Journal Article. Current: Journal Article|Randomized Controlled Trial|Comparative Study|Research Support, Non-U.S. Gov't. [Inspect PubMed record.](#)

PMID 36322060. Stored: Journal Article|Research Support, Non-U.S. Gov't|Comment. Current: Journal Article|Research Support, Non-U.S. Gov't. [Inspect PubMed record.](#)

PMID 40982552. Stored: Journal Article|Randomized Controlled Trial. Current: Journal Article|Randomized Controlled Trial|Research Support, Non-U.S. Gov't. [Inspect PubMed record.](#)

PMID 40286856. Stored: Journal Article|Randomized Controlled Trial. Current: Journal Article|Randomized Controlled Trial|Research Support, Non-U.S. Gov't. [Inspect PubMed record.](#)

These point-in-time differences are why stored publication types are treated as an evolving reference rather than immutable truth. PMID 41052730 changes from a stored tag-absent / assigned-RCT discordance to a current-tag-present / assigned-RCT match. This is one inspected case, not a population correction factor.

Comparison definitions

Positive-label concordance = $242,161 / 383,774 = 63.10\%$. Positive-tag coverage = $242,161 / 351,067 = 68.98\%$. Overall agreement = $(14,875,463 + 242,161) / 15,368,143 = 98.37\%$. The always-tag-absent baseline is $15,017,076 / 15,368,143 = 97.72\%$. These quantities describe tag-label agreement and must not be relabeled as independently validated clinical precision or sensitivity.

S5. Temporal lexical control and source audit

The diagnostic uses the released disease string and the following conservative regular expression. The cutoff is recorded publication years 2010-2018. It excludes generic coronavirus and SARS-CoV without “2”. Matching this literal pattern does not establish that the field contains a valid disease concept.

```
(^[^a-z])covid[- ]?19([^\0-9]|$)
OR sars[- ]?cov[- ]?2([^\0-9]|$)
```

Recorded year	Flagged records	Flagged documents
2010	7,258	5,880
2011	5,442	4,633
2012	5,965	5,027
2013	6,046	5,106
2014	6,538	5,478
2015	6,647	5,663
2016	6,792	5,674
2017	7,916	6,660
2018	7,631	6,447
Total	60,235	50,568

The denominator is 20,998,938 extraction rows dated 2010-2018. PMID linkage is present for 43,334 flagged rows. The 30-document source check is conditional on being flagged and PMID-linked; it is not a sample of the whole release. One row per distinct document is chosen, then documents are hash-ranked with the seed “HCC-temporal-audit-20260915”.

All 30 source records were retrieved; all contained an abstract. All 30 current journal years matched the stored recorded year. Zero current title/abstract pairs matched the diagnostic terms. Exact source identifiers and results are in temporal_source_audit.csv. Source abstracts are kept outside the distributable package; the CSV contains bibliographic metadata and check results.

Interpretation: a reliable historical analysis cannot treat these date-label combinations as contemporaneous mentions without additional evidence. The test does not establish the specific error mechanism or validate the other, unflagged disease labels. Retrospective updates, polluted input passages, incorrect linkage, and generative extraction errors require separate investigation.

S6. Independent annotation protocol - not yet performed

The supplied reviewer worksheet contains 294 extraction records from the 120-document publication-tag audit. Its review fields are intentionally blank. This is a starting set for rubric development, not a representative benchmark. A confirmatory evaluation needs a fresh probability sample and documented sampling weights.

Field to adjudicate	Reviewer instruction
Source/input identity	Confirm the exact input document and processing-time text; record unavailable when it cannot be recovered. A current PubMed abstract alone may differ from the model input.
Disease concept validity	Does the disease-field value represent a health condition in this context? Distinguish a named concept from a model-generated or non-disease string.
Factor concept validity	Is the factor identifiable in the supplied input? Record literal support separately from reasonable normalization.
Relation support	Does the input actually express the extracted link? Code supported / contradicted / not stated / unclear. Do not infer a relation from co-occurrence alone.
Direction and stance	Adjudicate these separately; unspecified is distinct from neutral or absent effect.
Study design	Classify the described study using a written human rubric. PubMed publication types are evidence to inspect, not automatic adjudication.
Input sufficiency	Would the truncated text available to the model be sufficient? Separate model error from missing context.

Use two independently working reviewers with relevant domain competence, blinded to each other's decisions. Resolve disagreements after retaining the original judgments. Report agreement, label-specific denominators, unavailable-input rates and an adjudication log. Estimate document-clustered uncertainty for relationship-level performance. Missing source text must not be silently scored as either correct or incorrect.

Do not report current audit percentages as gold-label performance. Clinical performance claims require a completed independent annotation study. The current preprint limits its claims to corpus measurements, input-level checks and explicitly defined lexical safeguards.

S7. Reproduction and provenance

The supplied analysis code reads the fixed HCC release, creates compact local analytical data, calculates the reported tables, constructs charts and builds the manuscripts. No live production database is required for the reported corpus results. Network access is used only for independently retrieved public bibliographic metadata. Those metadata are time-varying.

Core files

File / group	Purpose
analysis/PROTOCOL.md	Exploratory protocol and declared interpretive boundaries
analysis/ingest.py	Strict file ingestion and compressed-byte fingerprinting
analysis/analyze.py	Release-level aggregates, document comparisons and deterministic samples
analysis/derive.py	Fixed weights, sensitivity rates, decomposition and summary metrics
analysis/extensions.py	Balanced journal panel and temporal context counts
analysis/audit_pubmed.py; audit_temporal.py	Current public metadata and temporal lexical checks
analysis/charts.py; manuscript/build.py	Figures and review PDFs
results/*.csv; summary.json	Exact denominators and figure-source data
results/reviewer_worksheet.csv	Blank human-review fields; not an annotated benchmark
results/findings_ledger.csv	Claim-to-result traceability for manuscript review

Reproduction instructions and input acquisition details are in README.md. Analysis of the released archive is reproducible from those public files. Original model training, source selection and end-to-end extraction are not fully reproducible from the archive alone, because the release omits complete source snapshots, input spans and historical model hashes.

Recovered workflow evidence

Available workflow scripts specify a FLAN-T5-based extraction stage with greedy decoding and truncated text, and a separate study/evidence classifier distilled from model-generated labels. The current files document the design but are not proof of the exact code and checkpoint versions used for every historical row. A provenance ledger records inspected source-file hashes; infrastructure identifiers and credential-bearing source files are excluded from this review package.

Known documentation corrections are proposed separately: replace “verified study type” with “stored bibliographic publication type”; avoid treating generated fields as direct source self-reports; clarify partial-year coverage; distinguish pair-direction thresholds from scientific disagreement; and explain that numerical category codes are not validated evidence scores. HCC v1.0 is unchanged; this revision is a separate preprint and sample.

References

References [1]-[11] are in the main manuscript and manuscript/references.json. Audit CSVs include source links and retrieval times.

S8. Revision sample, recovery and denominators

Stratum	Frame docs	Sampled	Recovered	Assessed rows
Temporal holdout	35,248	300	299	397
COVID comparison	230,717	300	293	595
PMID early	6,767,320	100	100	217
PMID recent	8,139,607	100	98	224
Web early	2,744,920	100	96	212
Web recent	7,277,980	100	96	176

Discovery adds 30 sampled/recovered documents and 47 rows. Across all seven groups: 1,030 sampled documents, 1,012 uniquely recovered inputs, 1,896 sampled rows and 1,868 assessed rows. Eighteen inputs are missing. The 1,044 archive records include repeated identical inputs; no sampled document has conflicting archived inputs. These recovery failures are not silently coded as model errors or successes.

Holdout and comparison selections rank MD5(doc_id + HCC-revision-holdout-20260915). General strata rank MD5(doc_id + HCC-revision-general-20260915). Ties are broken by doc_id. All discovery cases are excluded; the general frames additionally exclude the already selected temporal/comparison cases. Strata represent deliberately different populations. Exact sample membership, frame sizes and archived-input hashes are in revision_sample.csv and revision_input_provenance.csv.

Definitions and intervention scope

The correction is an experimental selective-output layer. Original released rows are preserved in grounding_audit.csv; keep_joint_phrase is a new derived decision, and reason identifies unmatched entities or unavailable input. A missing/ambiguous input causes abstention. The retained derivative contains only assessed rows passing both phrase requirements, with original identifiers and an explicit non-validation warning. It is not a new clinically validated HCC version.

The original extraction client calculated a 60%-token-overlap diagnostic but inserted outputs regardless of that value. Our implementation makes rejection explicit and additionally checks full token-sequence boundaries. The comparison does not isolate the effect of each normalization choice or optimize a threshold. Negative controls and synthetic unit tests are disclosed separately from natural-source outcomes.

S9. Exact lexical coverage and mismatch controls

Stratum	Phrase kept / assessed	Legacy kept	Phrase retention 95% interval
Temporal holdout	13/397	15	3.3% (1.0-6.2)
COVID comparison	476/595	518	80.0% (75.0-84.2)
PMID early	182/217	210	83.9% (77.6-89.9)
PMID recent	173/224	205	77.2% (68.8-84.2)
Web early	183/212	198	86.3% (81.0-91.7)
Web recent	145/176	157	82.4% (73.5-89.8)

Intervals use 2,000 document-cluster bootstrap draws with seed 20260915. Whole document clusters are sampled with replacement within each recovered-input stratum, then retained rows are divided by total rows. This addresses within-document dependence for conditional coverage, not source selection, recovery bias or uncertainty about semantic truth. Discovery cases are excluded from the new-test comparisons.

General stratum	Mismatched rows	Phrase pass	Legacy pass
PMID early	217	0	0
PMID recent	224	0	0
Web early	212	0	0
Web recent	176	1	2

Each general stratum sorts its recovered document identifiers by MD5(doc_id + HCC-controls-20260915) and cyclically assigns the next document's text. No document maps to itself. Chance overlaps are retained and reported; the one phrase-pass case is not declared a confirmed semantic false positive. All constructed pairings are available in mismatched_input_controls.csv. These controls do not estimate real-world specificity.

Both entity phrases can be present while a relationship is negated or absent. Unit tests intentionally expect those examples to pass the lexical gate, exposing its scope rather than pretending it handles negation or entailment. Aliases and nonliteral paraphrases can be rejected. A human semantic benchmark is required to measure the resulting precision/recall trade-off.

S10. Replay provenance and public reproducibility limits

Replay stratum	Documents	Exact matches	Released rows	Replay rows
COVID comparison	20	20	40	40
PMID early	5	4	12	14
PMID recent	5	4	8	12
Web early	5	4	10	10
Web recent	5	3	15	17
Temporal holdout	30	27	44	51

The replay seed is HCC-replay-validation-20260915. This sample was selected before inspecting new generated outputs, after the recovery/lexical summaries. The three initial discovery replays are separate. Tuples include disease, factor, relation, epistemic stance and direction, with the documented length limits and release normalization. Tuple-list comparison preserves duplicate multiplicity. The model is not retrained for this experiment.

Retained model SHA-256: a0a096e61a6549770a6eaad6730c41ae856609680e8b3942a12d482ebd782a73

Config SHA-256: fb274a75f8e9620a02d3aaf66f82ac034a4d15aab662e6de030d667db5b48bdf

Greedy decoding; float16; maximum 768 input tokens and 256 generated tokens; one input per call. The retained checkpoint is not assumed to have a proven row-specific historical identity. Eight of 70 outputs differ from the released tuples; the public comparison CSV preserves every discrepancy.

The released-data analyses can be regenerated from the public HCC files. New input-level measurements can be recalculated from archived input text when access is available; hashes alone are insufficient to regenerate them. Full source passages and model weights are not included in this package. The gate, derived audit rows, sample identifiers, replay output strings, aggregate tables and figure sources are supplied. Public PubMed identifiers support source inspection, but current abstracts can differ from archived extraction inputs.

Human review remains pending. The blinded worksheet excludes model gate decisions and provides stable case identifiers. Two independent reviewers should retain their original judgments before adjudication; unavailable historical inputs must remain a distinct outcome. This preprint does not relabel its computational checks as such a completed review.