

Additional file 3 to: Can it be checked? A cross-sectional meta-research audit of access to, and reporting in, machine learning studies of pharmaceutical manufacturing and quality control.
Phani Kumar Balagam.

Protocol, deviations log and decision-rule log

File digests (SHA-256): PROTOCOL.md
fe4add0ec6df54b904d02de38385072c1bc6e0132bcaa2937d9b2c6d85911f31;
DEVIATIONS.md
bbcd1a768fec9d93f3f10e323486451dafac1cb3674a467f8def9c345461f83;
DECISION-RULE-LOG.md
ad0889244cd58de2fd202da7f2d463ee2ede64487faea8b599ee43109165e145. The protocol was fixed on 2026-09-23 before title-and-abstract screening and before any scoring, and has not been edited since.

PROTOCOL - Can It Be Checked? (submission #10, concept A13)

Version 1.0, fixed 2026-09-23 before title-and-abstract screening and before any scoring. The literature search had already run (13:55-14:05 UTC, same day), with search strings fixed in `harvest.js` (SHA-256 4d03e1255af25c2670a6248faa8cbc6fc1a674cb76dd207d1f951c8b4355115a) before it ran. Any change after this point goes in `DEVIATIONS.md` with date and reason; this file is not edited.

1. Question and design

A meta-research (methodological) cross-sectional audit of the peer-reviewed literature. Two outcomes, on two denominators, fixed in advance (author ruling D3):

- **Outcome 1 - access.** Among ALL eligible papers: is the full text readable by an unaffiliated reader, without a subscription?
- **Outcome 2 - reporting.** Among the READABLE eligible papers only: does the paper report enough for an independent reader to check its headline machine-learning result (items R1-R9 and a composite)?

Outcome 2 describes the openly readable literature only. The title, abstract and Limitations say so.

2. Eligibility (author ruling D1, "Q-narrow")

Included: peer-reviewed journal articles and full conference papers, **first published online** 2019-01-01 to 2025-12-31, that fit or evaluate at least one ML model on data generated by the **manufacture or quality control** of a human drug product, drug substance or biologic, and report at least one quantitative performance result. The full criteria (I1-I3) and exclusion codes (E1-E12) are Prompt 1 in `PROMPTS.md`, reproduced verbatim there.

Explicitly out: falsified-medicine market surveillance (E7), gene-therapy vector characterization (E8), point-of-care or hospital printing (E4), drug discovery (E3), clinical and pharmacovigilance work (E4), supply chain (E5), herbal or traditional medicine (E6), physicochemical property prediction without process or product data (E9), classical chemometrics alone (E1), reviews and preprints (E10). The author's own works are excluded by name.

First online date = Crossref published - online; if absent, the earliest date among Crossref published - print and issued; if Crossref has no record, the earliest date among the indexing sources.

3. Search

Five sources, one fixed query (three concept blocks: ML AND manufacturing/quality AND pharmaceutical anchor) on title and abstract: PubMed, Europe PMC, OpenAlex, Semantic Scholar, DOAJ (abstract field, split in four because of the URL limit). Harvest years 2018-2026; the date rule is applied afterward. Exact strings: `harvest.js` and Additional file (search strategy). Crossref and Unpaywall are used for metadata and access, not as search sources, because Crossref supports no boolean search (run 1: 1,769,185 hits for a bag-of-words query).

Deduplication: `merge.py` (DOI, PubMed ID, normalized title within one year). Automatic pre-screening removals: preprints and theses (preprint-server DOI prefixes), records whose every source date falls outside the window, and review types declared by PubMed or Europe PMC. A machine-assigned Semantic Scholar "Review" tag is NOT trusted and goes to screening.

4. Screening

Title and abstract screening by the model (Section 8) under Prompt 1. "Unclear" records go to full-text eligibility where a full text is readable. Where it is not, they are classified from title and abstract, and outcome 1 is reported both with and without them (sensitivity analysis). Full-text eligibility is re-checked at scoring (Prompt 2, Step 2).

5. Outcome 1 - readability rule (three sources, fixed in run 1)

For every eligible paper with a DOI, query Unpaywall, OpenAlex and Crossref.

- **Flagged open:** Unpaywall `is_oa` true, OR OpenAlex `is_oa` true, OR a Creative Commons license in the Crossref record.
- **Closed:** all three negative.
- **Readable (verified):** flagged open AND the full text was actually retrieved and contained the article body (Methods or equivalent). Every flagged-open paper is retrieved for scoring, so this is checked on every one.
- **Flagged open, not retrievable:** flagged open, but no copy with the article body could be retrieved (dead link, bot check, abstract-only page). Reported as its own category; counted as not readable in the primary analysis and as readable in a sensitivity analysis.
- **Undetermined:** no DOI, or all three services failed.

A paper counts as readable only when a copy with the article body was obtained. Bot checks and CAPTCHAs are never bypassed.

6. Outcome 2 - the rubric (author ruling D5)

Nine items, each traced to published standards. Checked against the official texts on 2026-09-23: REFORMS (Kapoor et al., Sci Adv 2024, doi 10.1126/sciadv.adk3452), DOME (Walsh et al., Nat Methods 2021, doi 10.1038/s41592-021-01205-4, Box 1), TRIPOD+AI (Collins et al., BMJ 2024, doi 10.1136/bmj-2023-078378), and the leakage taxonomy (Kapoor and Narayanan, Patterns 2023, doi 10.1016/j.patter.2023.100804).

Item	Short name	REFORMS	DOME (Box 1)	TRIPOD+AI
R1	Data availability	2a	Data: "Availability of data"	18e
R2	Code availability	2b	Model: "Availability of software"	18f
R3	Data split stated	5c	Data: "Data splits"	12c
R4	Held-out evaluation	5d, 6a	Data: "Redundancy between data splits"	12c
R5	Uncertainty on headline metric	7b	Evaluation: "Confidence"	23a
R6	Baseline comparison	5f	Evaluation: "Comparison"	-
R7	Class distribution (classification)	3d, 3e	Data: "Provenance"	-
R8	External validation	8a	Evaluation: "Evaluation method"	-
R9	Repeated evaluation	7b	Data section text on repeated splits by cross-validation or bootstrapping	-

R10 (ablation) is DROPPED. The run-1 draft mapped it to DOME's Model section. DOME Box 1, read in full on 2026-09-23, has no ablation or component-contribution question. Its Model questions cover interpretability, output type, execution time and software availability. Neither REFORMS nor TRIPOD+AI has an ablation item either. An item with no source standard cannot be kept (the rubric rule from run 1).

Composite "checkable" = R1 = Y and R2 = Y and R3 = Y and R4 = Y. That is the minimum an independent reader needs to re-derive the headline number: the data, the code, the partition and a held-out result.

Item values: Y, N, NA (not applicable; R7 for non-classification tasks), NASS (not assessable: the evidence exists but could not be reached). "Not reported" is N, never NASS. Decision rules are in Prompt 2, verbatim. Every Y, and every N that rests on a statement, records a locator and a verbatim quote of at most 15 words (author ruling D4).

Proportions are computed over papers scored Y or N for that item. NA and NASS are counted and reported but kept out of the denominator. The composite requires all four items to be assessable.

7. Descriptors

Year (first online date); venue type; task; ML family; data provenance.

Venue type is assigned from the journal or proceedings title into one of six classes, fixed here: - **pharmaceutical sciences**: the title names pharmacy, pharmaceuticals, pharmaceutical, drug delivery or drug development - **chemical and process engineering**: chemical, process or biochemical engineering, or computers in chemical engineering - **analytical chemistry and spectroscopy**: analytical, spectroscopy, chemometrics, talanta, separation or chromatography - **biotechnology and bioprocess**: biotechnology, bioprocess, biochemical or bioengineering, not already classed - **computing and engineering**: computing, AI, machine learning, electrical engineering, sensors, or IEEE/ACM proceedings - **multidisciplinary**: everything else, including mega-journals (for example Scientific Reports, PLOS ONE, Heliyon)

The first matching rule wins, in the order listed. The venue-to-class table is released.

8. Who scores, and the human check (author ruling D4)

Screening and scoring are done by a large language model, used as the scorer under this protocol. The model is Claude (Anthropic), model identifier `claude-opus-5-5`, run in the Claude desktop agent environment with parallel instances receiving the identical prompt. The author is the sole author and is accountable for every decision.

- **Prompts.** Both prompts are given verbatim in `PROMPTS.md` (an additional file).
- **Validity check.** A random 20% of scored papers (seed 20260924) goes into a blinded worksheet for the author. It carries the item definitions and the full-text links, and no model scores. When the author returns it, author-versus-model agreement is reported per item and for the composite: Cohen's kappa (Cohen 1960) and percent agreement. This is reported as a **validity check of the model scores, never as inter-rater reliability between two humans**.
- **Disagreements.** The author resolves every disagreement, and his score is the one analyzed for those papers.
- **If no worksheet comes back.** If the author does not return the worksheet, the paper ships with model-only scoring and says so plainly in Methods and Limitations.
- **Decision-rule log.** Every rule clarification made during scoring is logged with date and paper in `DECISION-RULE-LOG.md` and applied to every paper scored before it.

9. Analysis (fixed before scoring)

- Outcome 1: the proportion readable among eligible papers, with a Wilson 95% interval. Primary and sensitivity definitions are in Section 5.
- Outcome 2: per-item proportions of Y with Wilson 95% intervals, plus the composite.
- By venue type: per item, Fisher's exact test (2 x k; Freeman-Halton), with the Holm correction across the nine items plus the composite.
- By year: logistic regression of the item on year of first online publication (per-year odds ratio with 95% CI), with Holm correction across the same ten tests.
- Agreement statistics: as in Section 8.
- Null result: no year trend and no venue difference is a publishable descriptive baseline.

The paper is not worth publishing if fewer than 60 papers are readable and eligible (the stop rule from run 1). It is also not worth publishing if the composite's interval is wider than 25 percentage points.

10. Data and code release

Everything goes to Zenodo, under the author's action: - the raw harvest files and their digests - merged .jsonl, screening decisions and the access table - the score sheet with evidence locators and quotes of at most 15 words (no longer text from any article) - the decision-rule log and the prompts - all code

There is **no author-affiliation field** anywhere (author ruling D8). Only aggregates appear in the manuscript.

DEVIATIONS from PROTOCOL.md v1.0

Each entry: date, what changed, why, and effect. PROTOCOL.md itself is never edited.

D-01 2026-09-23 - operational wrapper around Prompt 1 (screening)

Each screening instance received Prompt 1 verbatim, preceded by this fixed operational wrapper, which carries file-handling instructions only and no screening rule:

OPERATIONAL WRAPPER (plumbing only; the screening rules are the prompt below, verbatim).

Read the whole batch file with the Read tool (use offset and limit to page through it if it is long; records are separated by lines beginning "### key:"). Screen every record. Write the output file with the Write tool as JSON Lines: exactly one JSON object per record, in the same order as the batch. Then verify with Bash (wc -l) that the output line count equals the record count in the batch header, and fix it if not. Do not use the web and do not read any other file. Your final reply must be one line only: "done <records> include=<n> unclear=<n> exclude=<n>".

Prompt 1 as delivered (placeholders substituted per batch) has SHA-256 8836c6e6102b9524c260bb4b9a41c2dace912cace38f28aabc6f03d9ba83ddda before substitution. 31 instances, 118 records each (3,658 records), outputs checked for record count and order: 31 of 31 exact.

D-02 2026-09-23 - full-text retrieval: operational changes (no change to the readability rule)

The rule in PROTOCOL.md Section 5 is unchanged. How copies were fetched changed in four ways:

- **Code v1.1.** `fetch_fulltext.js` v1.0 (SHA-256 b3c58710...668cf0) treated a file as PDF only from its content type. One repository (Zenodo) served a PDF as `application/octet-stream`, and v1.0 read it as HTML (a 3,008,126-character garbage record, discarded). One repository served only the author's Word manuscript. v1.1 (SHA-256 6fe9c996f3d263e289c2a68579385386d7403c4ad7911df5d209c7925dd1fe07) recognizes a PDF by its first bytes and converts `.docx` to text. Nothing else changed. Both versions are released.
- **How the code reached each page.** The code was carried between pages in the tab's `window.name` and evaluated on arrival, and its SHA-256 was checked on every page. Three kinds of site needed the code pasted directly into the page instead: sites whose Content-Security-Policy forbids `eval` (De Gruyter Brill); sites whose challenge page cleared `window.name` (PMC, RSC) or that were handled in a separate tab as a precaution (Taylor & Francis, IEEE, ACS); and single-PDF cases (MIT, ACM, WUR, 50sea, Journal of Neonatal Surgery). The pasted code was the same functions. On those pages only the PDF path of `grabOne` was used.
- **Promise.withResolvers.** Two DSpace 7 repositories (`ruc.udc.es`, `helda.helsinki.fi`) replace the browser's `Promise` with one lacking `Promise.withResolvers`, which `pdf.js` needs. A three-line shim was added on those two pages only.
- **Second hops.** Repository and journal landing pages were followed to the article file on the same host: DSpace `server/api/core/bitstreams/<id>/content`, OJS `article/download/<id>/<galley>`, Pensoft `article_preview.php`, J-STAGE and J. Braz. Chem. Soc. embedded PDFs, and IEEE `stampPDF`. Every copy came from a public URL. No bot check was answered or bypassed. A page that stayed on a challenge was recorded as `bot_check` (`RETRIEVAL-NOTES.csv`).

D-03 2026-09-23 - first-online date: rule applied literally, with a sensitivity check

The date rule in Section 2 was applied as written: 1. Crossref published-online; 2. otherwise the earlier of Crossref published-print and issued; 3. otherwise the indexing sources.

For many Elsevier records Crossref carries no published-online, so step 2 dated 15 articles by their 2026 issue. Those 15, plus 2 whose Crossref online date is in 2026, fall outside the window and are excluded as `E_date` (17 of 616). A sensitivity analysis dates each record by the earliest date in any source instead, and reports what changes.

D-04 2026-09-23 - one readable paper that the text pipeline cannot score

10.1016/j.ijpharm.2022.121949 is readable only as an image-only PDF: an 11-page accepted manuscript on HAL with no text layer. It counts as readable for outcome 1, because a reader can read it. Neither browser OCR nor a container download was possible (the host is proxy-blocked from both shells), so every rubric item is NASS for this paper. It is reported in Limitations.

D-05 2026-09-23 - which version of a paper counts as readable

The protocol does not restrict the version, so a readable copy may be the published version, an accepted manuscript or a submitted version: for example, an arXiv copy where the publisher copy was not retrievable. The version of each readable copy is recorded from Unpaywall's version field where available. Scoring uses the copy obtained, and the version is reported as a descriptor.

D-06 2026-09-23 - operational wrapper around Prompt 2 (scoring)

Each scoring instance got Prompt 2 verbatim, with only {BATCH_FILE} and {OUTPUT_FILE} filled in. Before it came this fixed operational wrapper: file handling and output shape only, with no scoring rule (score/wrapper.txt, SHA-256 82a5d2b3b6785a19eb50f83b24d14975eb8e01b08d0bbb630cf576e36278e510). After it came Prompt 1 verbatim, because Step 2 of Prompt 2 refers to the screening criteria. Prompt 2 as extracted from PROMPTS.md, before substitution, has SHA-256 051c4b7f49092ed710f503aec98069c4c8f3606acc0e2cbd29bcc99bc64f8f48.

Run: 39 instances with 6 articles each (the last had 3), 231 articles, in a seeded random order (seed 20260923). All 39 outputs matched their batch in count and order. Full texts were rewrapped at 1,000 characters per line so that no line exceeded the reader's limit. Text was not otherwise changed.

D-07 2026-09-23 - search notes written in the quote field

Prompt 2 asks that an N with no statement carry the evidence "not reported" and the search terms used. In 10 of the 186 scored papers, instances wrote that note into the quote field (65 fields). build_scoresheet.py moves any quote that starts with "not reported" into a separate search_note field. Values were not changed. Every other quote (824 first-pass quotes plus the 10 DR-01 quotes) was checked against its full text by validate_scores.py. All were found verbatim, after Unicode and whitespace normalization.

D-08 2026-09-23 - how the Fisher-Freeman-Halton p-value is computed

The protocol names Fisher's exact test (2 x k; Freeman-Halton). Exact enumeration over six venue classes was too slow for the build. The p-value is instead a Monte Carlo estimate of the same exact test: the statistic is the table probability under fixed margins, from 100,000 tables drawn with scipy.stats.random_table (seed 20260924), with the +1 correction. The standard error of a p-value near 0.05 is below 0.001.

D-09 2026-09-23 - venue classes applied literally

The six-class title rule in Section 7 was applied exactly as written, with no words added, and the venue-to-class table is released (`results/venue_classes.csv`). Specialist titles that lack the listed words fall into "multidisciplinary", for example "AAPS PharmSciTech", "Journal of Controlled Release", "Powder Technology" and "Organic Process Research & Development". That class holds 92 of 186 scored papers. This is reported as a limitation of the prespecified rule, and no reclassified analysis is substituted.

D-10 2026-09-23 - the title no longer carries the scope qualifier

Protocol Section 1 says that outcome 2 describes the openly readable literature only, and that the title, abstract and Limitations say so. By the author's ruling of 2026-09-23 (run 3), the title ends at "quality control"; the trailing clause "with reporting assessed in the openly readable literature" repeated what the abstract states. The abstract (Methods: "Outcome 2, among readable papers only") and the Limitations section still state the scope. No method, outcome, analysis or number changed.

D-11 2026-09-23 - a year model with no finite estimate is reported as not estimable

`analysis.py` already reported an item with no events, or no non-events, as not estimable (no odds ratio, $p = 1.0$ in the Holm family). A test of the validity-worksheet rebuild showed that an item with a single event can give a logistic fit with no finite odds ratio (quasi-separation), and the script then stopped. Such a fit is now reported the same way as the no-event case. This applies to no result in the manuscript: re-running the analysis on the scores reported here reproduces every result file byte for byte.

D-12 2026-09-24 - how the 20% validity check was run, and what agreement is computed over

Protocol Section 8 planned a blinded worksheet in which the author scores every item of the 20% sample, with author-versus-model agreement reported per item and for the composite. That check was prepared and not completed: no person rescored any paper. The unfilled worksheets are released in Additional file 6. What ran instead, on 2026-09-24: - Every paper in the sample (37 of 185, seed 20260924) was scored again by a second model pass: Claude (Anthropic), model identifier "claude-opus-5", in a session separate from the first pass. It wrote its scores before any first-pass output was opened. - Every item where the two passes disagreed was re-read against the verbatim rule, scoped to the paper's headline result, and resolved with a recorded one-line reason. A random 25% of the agreeing items (seed 20260924) was re-read the same way. The remaining agreeing items were accepted without re-reading. The resolution was model work. - Each item cell carries its provenance: adjudicated (the passes disagreed), reviewed (agreeing, re-read) or accepted (agreeing, not re-read). The record is released in Additional file 6.

Consequences for the analysis: - The reported validity statistic is the blind two-pass agreement: percent agreement and Cohen's kappa between the first and second passes, per item and overall, over all item cells of the sample. It measures the reproducibility of the rubric under the same model. It is not inter-rater reliability and not accuracy against human judgment. No agreement is computed between the resolved values and the first pass, since the resolved values were chosen from the two passes. - The resolved value replaces the first-pass score wherever

the two differ, and the analysis runs on the resolved sheet. - The first-pass (model-only) results are reported alongside as a sensitivity analysis. The protocol did not pre-specify this sensitivity analysis; it is added here.

Reason: the author's decision of 2026-09-24 to complete a validity check before submission. The blinded author check was prepared and not completed, and this is the procedure that was run.

DECISION-RULE LOG (PROTOCOL.md Section 8)

Every clarification of a decision rule that was made during scoring. Each entry was applied to every paper scored before it: the papers it could change were re-scored under it, and the rest are unaffected by construction.

DR-01 2026-09-23 - R1, journal boilerplate data statements

Trigger. The first-pass validation of all 185 eligible score records (`validate_scores.py`, then a read of every R1 = Y quote) found 10 papers whose only R1 evidence was a journal-template statement: "All relevant data are within the manuscript and its Supporting Information files", "All data generated or analyzed during this study are included in this published article", "The original contributions presented in the study are included in the article/Supplementary Material", or "The data used is publicly available" with no location. A template statement is not a location.

Clarification. Under R1, a data-availability statement counts as Y only when it locates the data behind the headline result: - a named repository, DOI or URL; - a named supplementary file that the statement or the text says holds the data; or - a table or appendix in the article that holds the modeled observations, found in the text.

If the statement points only to supplementary files, and nothing in the text shows what they hold, the value is NASS (the evidence is in a part of the article that could not be reached). If the data are printed in the article, give the table. Otherwise the value is N.

(The instruction given to the model instances spelled "modeled" with a double l; it is reproduced as given in Additional file 4.)

Applied to. The 10 papers with a template-only R1 = Y quote (listed in `score/dr01_batch.txt`) were re-scored on R1 alone, by one model instance per 5 papers. Every other R1 value is either N, which a stricter rule cannot change, or Y resting on a located repository, file or table.