Research in Digital Experimentation
Vol. 1 No. 1 (2026)

A Searchable Disclosure Can Be Missing from Archive Front Matter

Frank Mills · Claire Harris · James F Harrison

Published: November 26, 2026 | Article 1224

Abstract

PubMed search evaluates the presence of a disclosure field. An importer working from PubMed Central needs a separately typed disclosure element in the downloaded front matter. We tested three administrative signals on 120 exact DOI pairs. PubMed's disclosure search and citation XML agreed on all records, identifying 114. The selected PMC front-note path exposed 38, all among those same 114. A general author-notes container was present on 68 fronts and would have admitted 30 records without the typed note. Every typed note had a local identifier, while none had a direct contributor reference to that identifier. Consistent search retrieval, portable markup and contributor attribution therefore had different coverage on the same publications.

A Searchable Disclosure Can Be Missing from Archive Front Matter

Abstract

PubMed search evaluates the presence of a disclosure field. An importer working from PubMed Central needs a separately typed disclosure element in the downloaded front matter. We tested three administrative signals on 120 exact DOI pairs. PubMed's disclosure search and citation XML agreed on all records, identifying 114. The selected PMC front-note path exposed 38, all among those same 114. A general author-notes container was present on 68 fronts and would have admitted 30 records without the typed note. Every typed note had a local identifier, while none had a direct contributor reference to that identifier. Consistent search retrieval, portable markup and contributor attribution therefore had different coverage on the same publications.

1. A searchable field answers a particular question

A disclosure can move through an editorial system, a published record, and bibliographic services that each expose different elements. To find publications for which PubMed has a disclosure field, a user can apply the official search selector. To extract typed notes from a front-matter archive, a user needs to find the appropriate element in that archive format. To associate a note with a named contributor position, a field extractor needs an explicit link between the contributor and the note. These tasks can begin from the same DOI. An observation from one path gives limited evidence about the others unless those paths are checked on the same records.

Guidelines address why publishers ask for such reporting. The International Committee of Medical Journal Editors (n.d.) discusses disclosure of financial and non-financial relationships and activities. Taichman et al. (2020) describe an electronic disclosure form intended to harmonize reporting arrangements. Blum et al. (2009) document variation in the requirements stated by journals, and Grundy et al. (2018) examine how published disclosures can be studied as reporting evidence. A policy about collecting or presenting declarations does not select a single archive position for them. The public fields compared here are useful for studying discoverability. They say nothing about whether an individual author has a particular relationship or whether a declaration is substantively complete.

A more specific bibliographic distinction is visible in the available documentation. The National Library of Medicine (2025) defines the CoiStatement element in the PubMed citation schema. Its user guide describes a search filter, hascois, for records carrying this kind of statement (National Library of Medicine, n.d.). PMC provides an openly available front-matter feed for structured archive metadata (PubMed Central, n.d.-a). Its tagging guidance describes more than one placement for this sort of material in article XML (PubMed Central, n.d.-b). Our test uses only one expressly defined front path. The archive and citation services are related and can receive common inputs. Each interface exposes a specified portion of the bibliographic record.

A separate example illustrates the scope of the signals. PubMed and PMC both identify Yi et al. (2024), DOI 10.1016/j.esmoop.2024.102239, through agreeing bibliographic keys. The exact PubMed citation has no searchable disclosure marker in the measured field. The archive front contains an author-notes container and no direct typed disclosure note in that container. We do not examine other article locations or semantic content. The example demonstrates why a generic container does not establish the more specific markup even when the work match itself is secure.

The research questions follow these levels of evidence. First, does the hascois query return the same citations whose XML exposes a nonempty disclosure element? Second, how often does a matching PMC front expose a recognized typed disclosure note, and does a general author-notes container approximate that availability? Third, where a typed note is observed, how often is it assigned a local identifier and how often does a contributor element reference that identifier? These comparisons measure usable indicators for indexing and extraction under the specified public paths.

2. Administrative disclosure data as a measurement problem

A body of empirical work has measured how visible reporting fields are. Iqbal et al. (2016) examine several indicators of reporting and transparency in a sample of published research. Serghiou et al. (2021) describe rules for extracting transparency indicators from article XML, including examples where a published format or its accessible markup matters. Rogus et al. (2024) study the adoption of the voluntary disclosure statement displayed through PubMed. Their work concerns variation in posting practices over time and coverage across journals. The present design asks about interchangeability at a fixed time and identity. It observes the searchable signal and source element together, then tests a narrower archive path for portability and explicit note-to-contributor links. It does not infer presence elsewhere in the article.

The visible structure can affect downstream quantitative work. Delgado-Quirós and Ortega (2024) distinguish the coverage of publications from the completeness of their metadata across free research databases. Visser et al. (2021) assess record and citation coverage, while Culbert et al. (2025) show that citation reference coverage can vary even for work records already represented. Hendricks et al. (2020) describe how distributed metadata deposits underpin scholarly infrastructure. De Jonge and Kramer (2026) examine how field completeness can relate to production workflows. Their observations motivate holding publication identifiers constant and checking the actual paths that an application would read. A successful bibliographic join establishes the common publication. It does not harmonize the metadata trees around that key.

JATS provides a concrete illustration of the difference between tagging a note and assigning it to a contributor. According to the National Information Standards Organization (2024), the footnote type attribute can distinguish a disclosure from other notes, with an older conflict value and the more specific coi-statement form. JATS4R (n.d.) gives separate markup examples for a general note and a note referenced by contributor elements. PLOS (n.d.) offers a concrete journal policy on providing competing-interest information. That policy and the XML cross-reference mechanism sit at different stages of the process. Presence of a typed, locally identifiable note supports document-level discovery before any claim about its applicability to a named contributor can be encoded or checked.

3. Sampling and observation

3.1. Establishing paired publication identity

The six source strata were 2024 PubMed ISSN search frames for PNAS, PLOS Biology, PLOS ONE, Nature Communications, Scientific Reports and eLife. We reused the archived first 100 records sorted by publication date for each query. From those archived candidate lists we required a PubMed ID, one DOI and a PMC accession. Each qualifying frame was ranked by a deterministic SHA-256 score of the DOI using the fixed prefix in the acquisition script. We took the first 20 pairs per journal, keeping all six strata equal in size. These criteria give a useful finite test of corresponding records. They do not constitute a census of each title or of all PMC articles from 2024.

We retrieved PubMed citation XML and PMC front-matter XML separately. PubMed's APIs and their role in exchanging bibliographic metadata are described by Sayers et al. (2024). The International DOI

Foundation (n.d.) describes the DOI as a maintained identifier for a registered object independently of the representation used to access it. We required agreement on the expected DOI in both retrieved records after trimming and case-folding. PMC's record, when it supplied an explicit PMID, also had to agree with the sampled PMID. Matching did not rely on a title, author name or string similarity. Christen (2012) distinguishes matching observations by their identity keys from indirect record linkage. The exact checks establish a common work before testing whether the two views expose the same evidence.

3.2. Two visibility checks and one targeting check

The PubMed citation indicator is one when the CoiStatement element contains a nonempty payload. The paired search indicator is one when the explicit PMID appears in a hascois query restricted to the requested identifiers. The search and citation XML were requested separately in the same acquisition window. We checked returned PMID sets before comparing their flags. The archive indicator is one when the retrieved PMC front's article metadata has a nonempty, direct footnote in author-notes with the type coi-statement or conflict. We also counted general author-notes containers and the number of direct footnotes per record. Other XML locations are outside the requested path, so a negative value represents no qualifying node at this location in the measured front. It supplies no evidence about whether other locations contain a declaration.

To test contributor-level targeting, we saved the count of recognized nonempty notes with a local node identifier and checked for exact references from each direct contributor's cross-reference element. XML makes local identifier values suitable for verifying an explicit reference, as formalized by the World Wide Web Consortium (2008). We match the target against the identifiers of the already recognized notes. Contributor positions and note counts are retained only as structural evidence. The observed data do not include contributor names, whole author lists, statement text, relationships described in a statement or scientific content. The difference between having a local identifier and being a target of a contributor reference is testable using only the document structure.

The primary summaries are paired counts by journal and aggregate across the 120 matched works. Within the archive front, we distinguish the existence of a general container from a typed note and an explicit pointer. This follows the emphasis of Park and Tosaka (2010) on semantics as a component of metadata quality. Wang and Strong (1996) frame quality around a user's intended application, and Batini et al. (2009) compare ways to operationalize data quality dimensions. In this study the definition of availability is tied to the search and XML field required for the particular task. No inference about disclosures from an individual or a publisher is warranted by an indicator's absence.

4. Results

4.1. Search and citation XML agreed across the matched frame

All 120 sampled records satisfied exact bibliographic identity. The PubMed hascois filter returned 114 citations and the same 114 had nonempty CoiStatement elements in the downloaded citation XML. The remaining six were negative under both observations. This complete agreement shows that the search field and visible XML payload worked interchangeably for selecting the observed PubMed records at this snapshot. It holds within the particular frozen frame, acquisition and query restrictions used here.

The recognized PMC front-note path returned 38 positives, all of which were also positive in PubMed. Seventy-six pairs carried the citation-level indicator alone, and six showed neither of the selected fields. The fraction of PubMed-positive records with a typed note at the measured PMC location was therefore 38 of 114, or one third. This is a conditional overlap for the fixed frame. Its denominator is restricted to exact DOI pairs with a positive PubMed field.

Table 1. *Disclosure metadata visible on 120 exact PubMed–PMC publication pairs.*

Journal	DOI-aligned records	PubMed field and search positive	Author-notes container	Typed PMC note	PubMed only	Neither measured field
PNAS	20	18	18	0	18	2
PLOS Biology	20	19	19	19	0	1
PLOS ONE	20	19	19	19	0	1
Nature Communications	20	20	0	0	20	0
Scientific Reports	20	18	0	0	18	2
eLife	20	20	12	0	20	0
All six	120	114	68	38	76	6

Note. The PubMed column shows identical membership under the search and citation XML rules. A PMC note qualifies if it occurs in the defined front path with a recognized disclosure type and nonempty payload. No qualifying front note appeared on a paired record without the PubMed citation-level marker.

The pattern of typed notes differed by journal in the selected sample. Nineteen records in each of the two PLOS strata displayed a qualifying note. The four other groups supplied between 18 and 20 citation-level positives and none with the recognized typed front note. In PNAS and eLife, some fronts still declared an author-notes container. In the two other groups, no direct container was present under the examined article-metadata path. Both observations are compatible with an alternative XML placement, a separate article representation or another provider process. The study contains the projected fields needed to test the counted paths and no evidence for deciding among those explanations. The stratum totals make this concentration explicit, rather than treating the overall 38 count as typical across journals.

4.2. A typed, locally identified note supplied no direct contributor link

Sixty-eight matched fronts had a general author-notes container. They contained 82 direct footnote nodes in total. Thirty-eight of those nodes qualified as typed nonempty disclosures. All 38 used the current coi-statement value after case-folding. None used the legacy conflict value in the observed front projection. A reader that treated the presence of an author-notes container alone as sufficient would identify 68 works, including 30 without a qualifying note. Since containers can hold other kinds of information, this rule would not recover the same administrative predicate.

Every qualifying note carried a local identifier. The contributor cross-reference checks found zero direct links to any of those 38 identifiers. A note ID by itself can make a string addressable within the document tree, but it is not a source observation that a given contributor's node targeted the note. General notes are represented possibilities in the markup guidance. These results should therefore be read as limited contributor-level addressability through one exact XML route. We do not interpret whether any observed note was intended to apply to an individual, to the whole group, or to a particular subset.

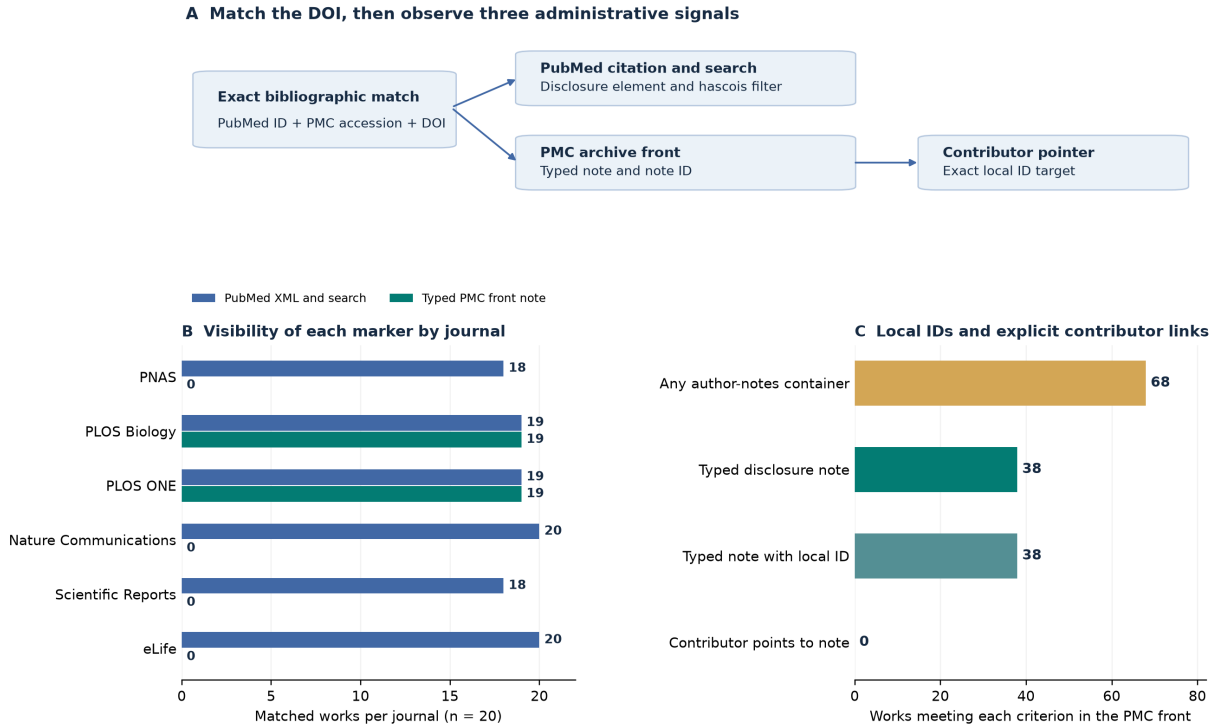

Figure 1. *Panel A shows the identifier match and the different paths through which metadata can be recognized. Panel B gives the observed frequency of each indicator within the six equal journal frames. Panel C compares an author-notes container, a typed note, a typed note's local ID and a direct contributor pointer in the captured archive front.*

5. Which evidence to preserve in an export

The search-versus-XML results identify an effective route for a specific visibility check. Someone building a dataset of citations with a nonempty published PubMed disclosure field can use that search flag as a first selection and test it against the paired citation XML. Someone requiring a portable typed note from front matter has a smaller observed set available. An extraction tool that treats those datasets as interchangeable would silently place many records in the wrong retrieval frame. In this sample 76 works have the citation-level marker while lacking the selected typed archive path. The divergence is observable despite exact publication identifiers and a current public response from both services.

A narrower problem appears in contributor attribution. Exporting a note with its local identifier preserves the identity of the node through serialization within a document. An explicit contributor pointer would add a separate structural observation. If source metadata supplies a general note, a reader can preserve it at work level. If it supplies explicit cross-references, a parser can retain those relationships without interpreting the note itself. This allows discovery and representation to remain testable independently of the truth, scope or relevance of a disclosure. The paper consequently makes no estimate about undisclosed interests or differences in underlying compliance between publishers.

The comparison inherits constraints from its design. Reusing first-page journal frames gives manageable samples with verified programmatic origins and no claim to broad journal coverage or historical trends. Changes in data deposit, PMC front packaging or issuer practices may change the observed frequency in a later acquisition. The schema's valid full-article placements also give reasons why a front-only scope can miss a separately published record of a statement. The purpose here is to quantify what one selected path actually supplies on exact matched citations, which can be independently replayed. Wilkinson et al. (2016) emphasize reusable metadata, Peng (2011) describes the distinctions between replay and independent replication, and Gebru et al. (2021) identify the value of documenting dataset selection. The

archived identifiers and compact field flags let a future user reproduce the current counts and choose a different structural test while keeping the change in method explicit.

Data and code availability

The reproduction record identifies the frozen administrative source projections, search receipts, scripts and figure. Author contributions, funding and competing-interest declarations await author confirmation.

References

Batini, Carlo; Cappiello, Cinzia; Francalanci, Chiara; Maurino, Andrea (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41(3), 1–52*. DOI.

Blum, Jared A. (2009). Requirements and Definitions in Conflict of Interest Policies of Medical Journals. *JAMA, 302(20), 2230*. DOI.

Christen, Peter (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. *Springer Berlin Heidelberg*. DOI.

Yi, L.; Ning, Z.; Xu, L.; Shen, Y.; Zhu, X.; Yu, W.; Xie, J.; Meng, Z. (2024). The combination treatment of oncolytic adenovirus H101 with nivolumab for refractory advanced hepatocellular carcinoma: an open-label, single-arm, pilot study. *ESMO Open, 9(2), 102239*. DOI.

Culbert, Jack H.; Hobert, Anne; Jahn, Najko; Haupka, Nick; Schmidt, Marion; Donner, Paul; Mayr, Philipp (2025). Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics, 130(4), 2475–2492*. DOI.

de Jonge, Hans; Kramer, Bianca (2026). Manuscript submission systems and metadata completeness in Crossref: Patterns and associations. *PLOS One, 21(3), e0345417*. DOI.

Delgado-Quirós, Lorena; Ortega, José Luis (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies, 5(1), 31–49*. DOI.

Gebru, Timnit; Morgenstern, Jamie; Vecchione, Briana; Vaughan, Jennifer Wortman; Wallach, Hanna; III, Hal Daumé; Crawford, Kate (2021). Datasheets for datasets. *Communications of the ACM, 64(12), 86–92*. DOI.

Grundy, Quinn; Dunn, Adam G.; Bourgeois, Florence T.; Coiera, Enrico; Bero, Lisa (2018). Prevalence of Disclosed Conflicts of Interest in Biomedical Research and Associations With Journal Impact Factors and Altmetric Scores. *JAMA, 319(4), 408*. DOI.

Hendricks, Ginny; Tkaczyk, Dominika; Lin, Jennifer; Feeney, Patricia (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies, 1(1), 414–427*. DOI.

International Committee of Medical Journal Editors (n.d.). Disclosure of Financial and Non-Financial Relationships and Activities, and Conflicts of Interest. *ICMJE Recommendations*. Source.

International DOI Foundation (n.d.). DOI Handbook. *DOI Foundation*. Source.

Iqbal, Shareen A.; Wallach, Joshua D.; Khoury, Muin J.; Schully, Sheri D.; Ioannidis, John P. A. (2016). Reproducible Research Practices and Transparency across the Biomedical Literature. *PLOS Biology, 14(1), e1002333*. DOI.

JATS4R (n.d.). Conflict-of-interest (COI) statements. *JATS for Reuse*. Source.

National Information Standards Organization (2024). Type of Footnote. *Journal Publishing Tag Library, JATS version 1.4*. Source.

National Library of Medicine (2025). CoiStatement element. *PubMed XML Element Descriptions*. Source.

National Library of Medicine (n.d.). PubMed User Guide. *PubMed Help*. Source.

Park, Jung-Ran; Tosaka, Yuji (2010). Metadata Quality Control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms. *Cataloging & Classification Quarterly, 48(8), 696–715*. DOI.

Peng, Roger D. (2011). Reproducible Research in Computational Science. *Science, 334(6060), 1226–1227*. DOI.

PLOS (n.d.). Competing Interests. *PLOS ONE Publication Policies*. Source.

PubMed Central (n.d.-a). PMC Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) API. *PubMed Central Documentation*. Source.

PubMed Central (n.d.-b). PMC Tagging Guidelines: Document Objects. *PubMed Central Documentation*. Source.

Rogus, Stephanie; Ross, Joseph S.; Lurie, Peter (2024). Adoption of the voluntary conflict of interest statement on PubMed. *PLOS ONE, 19(10), e0308782*. DOI.

Sayers, Eric W; Beck, Jeff; Bolton, Evan E; Brister, J Rodney; Chan, Jessica; Comeau, Donald C; Connor, Ryan; DiCuccio, Michael; Farrell, Catherine M; Feldgarden, Michael; Fine, Anna M; Funk, Kathryn; Hatcher, Eneida; Hoeppner, Marilu; Kane, Megan; Kannan, Sivakumar; Katz, Kenneth S; Kelly, Christopher; Klimke, William; Kim, Sunghwan; Kimchi, Avi; Landrum, Melissa; Lathrop, Stacy; Lu, Zhiyong; Malheiro, Adriana; Marchler-Bauer, Aron; Murphy, Terence D; Phan, Lon; Prasad, Arjun B; Pujar, Shashikant; Sawyer, Amanda; Schmieder, Erin; Schneider, Valerie A; Schoch, Conrad L; Sharma, Shobha; Thibaud-Nissen, Françoise; Trawick, Barton W; Venkatapathi, Thilakam; Wang, Jiyao; Pruitt, Kim D; Sherry, Stephen T (2024). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 52(D1), D33–D43*. DOI.

Serghiou, Stylianos; Contopoulos-Ioannidis, Despina G.; Boyack, Kevin W.; Riedel, Nico; Wallach, Joshua D.; Ioannidis, John P. A. (2021). Assessment of transparency indicators across the biomedical literature: How open is open? *PLOS Biology, 19(3), e3001107*. DOI.

Taichman, Darren B.; Backus, Joyce; Baethge, Christopher; Bauchner, Howard; Flanagin, Annette; Florenzano, Fernando; Frizelle, Frank A.; Godlee, Fiona; Gollogly, Laragh; Haileamlak, Abraham; Hong, Sung-Tae; Horton, Richard; James, Astrid; Laine, Christine; Miller, Pamela W.; Pinborg, Anja; Rubin, Eric J.; Sahni, Peush (2020). A Disclosure Form for Work Submitted to Medical Journals: A Proposal From the International Committee of Medical Journal Editors. *JAMA, 323(11), 1050*. DOI.

Visser, Martijn; van Eck, Nees Jan; Waltman, Ludo (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies, 2(1), 20–41*. DOI.

Wang, Richard Y.; Strong, Diane M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems, 12(4), 5–33*. DOI.

Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; 't Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3(1), 160018*. DOI.

World Wide Web Consortium (2008). Extensible Markup Language (XML) 1.0, Fifth Edition. *W3C Recommendation*. Source.