Studies in Efficient Intelligence
Vol. 1 No. 1 (2026)

Testing the Year Used to Measure a Reference's Age

Ronald Greene · Cynthia Murray · James F Harrison

Published: November 26, 2026 | Article 1243

Abstract

For 2,008 deposited references with comparable years, 124 cited years differed from the year on the corresponding Crossref DOI record. In 109 of those cases the reference year matched an explicit print or online year in the same record. Changing the year source moved 13 positions across a two-year recency threshold and nine across a five-year threshold. The pooled median age stayed at seven years. A separate 578 positions lacked a structured reference year but had a registered DOI year. The choice of date source thus changes which references can be analyzed more often than it changes classifications among references with both years.

Testing the Year Used to Measure a Reference's Age

Abstract

For 2,008 deposited references with comparable years, 124 cited years differed from the year on the corresponding Crossref DOI record. In 109 of those cases the reference year matched an explicit print or online year in the same record. Changing the year source moved 13 positions across a two-year recency threshold and nine across a five-year threshold. The pooled median age stayed at seven years. A separate 578 positions lacked a structured reference year but had a registered DOI year. The choice of date source thus changes which references can be analyzed more often than it changes classifications among references with both years.

One reference and two observed years

The age of a cited work is often computed by subtracting the work's publication year from the year of the citing article. In a registry, a reference position can include its own structured year as well as a DOI that points to a separate bibliographic record. The two sources of the year can disagree even when the DOI is exactly the same. The difference may change the estimated age by one year or move the reference across the boundary of a recency indicator such as publication within five years. A direct comparison can be performed without reconsidering the substantive claims of the works involved.

Studies of the aging of references have used medians, recent-reference shares and distributional models to describe how styles of citation differ across fields and periods (Marx and Bornmann, 2016) (Price, 1965) (Lyu et al., 2025). Parametric studies have modeled the shape of age distributions instead of reducing a list to one measure (BARNETT et al., 1989) (Stacey, 2020). Other work compares the age of cited literature over long windows (Egghe, 2010) (Chow et al., 2023). In each case the result depends on a rule assigning a year to the cited publication. Our question is measurement-based. If the cited year from the reference record and the cited work's registered year can be paired on a given DOI, how often does changing between them alter age estimates or recency membership?

Publication dates can refer to an online release, issue assignment or another stage of circulation. Studies of metadata in bibliographic indices have shown that the stages are distinguished imperfectly by a single imported date (Das and Das, 2006) (Liu et al., 2026) (Zhu, 2025a, 2025b) (Haustein et al., 2015). Work measuring the effect of publication timing on citation counts further illustrates the consequences of seemingly small timing decisions (Tort et al., 2012) (Donner, 2018). We hold the citing work and the reference position fixed. That leaves a paired choice between the already-exposed reference year and the registered publication year of the exact target. We use the target's separately typed dates to inspect whether an alternative calendar role can explain a year difference descriptively. A match to a different role is recorded as compatibility with the available fields, not evidence of which event the reference compiler intended.

Construction of a paired administrative sample

The original public sampling frames contain all Crossref DOI rows selected by four known journal ISSNs in December 2024. We use Signal Transduction and Targeted Therapy, Nature, PNAS and JAMA. An existing frozen roster supplies the available DOI pool. For each journal, a fixed hash order of the normalized DOI string determines 12 source works. This selection occurs before the presence of reference years or resolved DOI links is examined. The known target work for the administrative illustration is held out of this main comparison and fetched independently. The panel is designed to compare encoding choices on the same set of citation positions. It does not estimate a population prevalence of reference-date disagreement across all publishing.

Crossref supports entry-level years as part of a deposited bibliography and accepts exact DOI links, alongside less structured citation forms (Crossref, n.d.-a) (Crossref, n.d.-b). Metadata conventions elsewhere also provide dedicated publication or issued fields for reference dates (National Information Standards Organization, 2021) (Citation Style Language, n.d.). Studies of reference structure show that the metadata needed to identify a cited work may be distributed across components or absent in some records (dos Santos et al., 2022). We read only the already structured year and DOI for each deposited reference position. We do not search prose for missing years. A legitimate bibliography can therefore contain dates that we classify as unavailable under this extraction rule. The source scope is what the public response actually exposes as separate fields.

For every distinct reference DOI, we query the corresponding Crossref record using an exact DOI filter. We normalize case and require returned DOI identities to be part of the requested set. References without a returned exact hit remain as unresolved DOI links. This avoids combining year disagreement with errors introduced by fuzzy title or name matching (Fellegi and Sunter, 1969) (Christen, 2012). Bibliographic index comparisons identify source coverage and field completeness as distinct dimensions (Visser et al., 2021) (Delgado-Quirós and Ortega, 2024) (Delgado-Quirós et al., 2023). Research on malformed reference DOIs also suggests why it is useful to retain unresolved identifiers as their own observations (Cioffi et al., 2022). The methods here keep the relevant DOI and year fields and response receipts. No bylines, titles, reference texts, abstracts or substantive content are stored.

Paired age and threshold comparisons

For a complete pair, the citing work has a valid year in the selected registry field, the deposited reference has a four-digit year, and an exact DOI response for the cited work provides a valid year. We use a supported range from 1500 through 2099 and treat anything outside it or containing additional non-whitespace characters as unresolved for this analysis. We compute two ages from the same citing year: one subtracts the declared cited year, and the other subtracts the general year in the cited work record. A positive difference between the latter registered year and the former declared year means the registered value makes the reference younger. We preserve the sign to distinguish changes in the two directions.

The primary comparison counts paired positions with any difference between the two years. A second comparison tallies the references that move into or out of a recency window of two years or five years, including the citing year as age zero. The windows use nonnegative elapsed calendar years. The samples count reference positions, so two different papers citing the same DOI contribute separate observations. The two age measures use the same denominator for every paired table. A structured year that is missing from a reference position can sometimes be supplied from a linked DOI record, but such a position would enlarge the denominator. We report that potential coverage gain separately from the paired sensitivity.

We also compare the declared year to the explicit online and print years for positions where the unqualified registered year differs. If either observed role equals the declared year, the discrepancy is consistent with the representation of a single work at different calendar stages. The remaining disagreements have no such corroboration in the returned fields. More detail about the metadata workflow would be required to attribute their causes. Publisher deposit coverage varies across administrative fields (de Jonge and Kramer, 2026). Following computational replay practices, the summary therefore preserves exact field provenance and the different roles used in the comparisons (Wilkinson et al., 2016) (Sandve et al., 2013) (Mesirov, 2010).

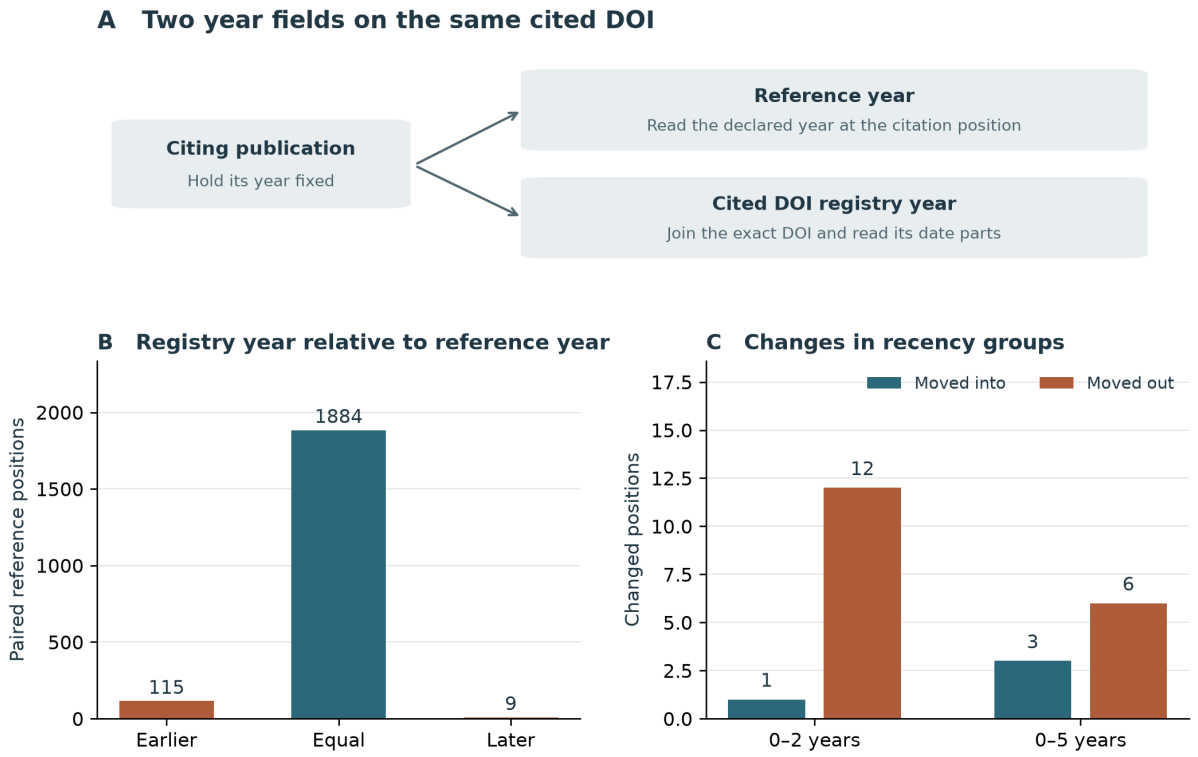

Figure 1. The study holds a cited DOI and citing year fixed and changes the source of the cited year. The quantitative panels show the direction of the year difference and the movements across two recency definitions.

What the paired years show

The 48 panel works have 2,777 deposited reference positions. Thirty-six works have a nonempty reference list in the returned metadata. Of the positions, 2,119 supply a valid structured year, 2,589 supply a DOI and 2,588 have an exact DOI response. A general publication year is available on 2,586 of those resolved positions. Every citing work carries 2024 in its selected date. The 2,008 positions that support both age calculations therefore hold the citing year and each reference position constant. We do not add the 578 DOI-resolved entries that lack a stated reference year to those paired comparisons.

Table 1. Each row counts positions in the citation lists of 12 works. The thresholds are inclusive. A changed reference moves into or out of the indicated recency group under the registered-year definition.

Citing journal	Reference positions	With a stated reference year	Paired on the exact DOI	Paired year disagreements	Change in the 0–2 year window	Change in the 0–5 year window
Signal Transduction and Targeted Therapy	1,944	1,880	1,789	114	10	8
Nature	47	45	44	4	1	1
PNAS	603	18	1	0	0	0
JAMA	183	176	174	6	2	0
Total	2,777	2,119	2,008	124	13	9

The DOI record's general year agrees on 1,884 of the paired positions. Among the remaining 124, the registered year is earlier than the declared reference year on 115 and later on nine. Most differences are a single year. The registered year is one year earlier on 113 positions and one year later on seven. Two positions are two years earlier. Two others have later differences of 17 and 32 years. These unusual cases are left in the summary and are not interpreted as bibliographic mistakes without independent evidence. Of the 124 differing positions, 109 have a reference year that matches the separately stated online or print year on the exact DOI. On the other 15 the available typed fields do not reproduce the declared year.

The effect is not balanced on the smaller recency threshold. Thirteen positions change at the two-year boundary, with one moving into and twelve moving out of the recent category. The count decreases from 410 to 399. The five-year definition changes nine positions, with three moving into and six moving out, reducing the recent count from 801 to 798. The corresponding net decreases are 0.55 and 0.15 percentage points of the compared positions. A comparison of the aggregate percentages alone would hide some movements because a reference entering the group can cancel the effect of another reference leaving it. The median in the pooled paired group stays at seven years. Of the 25 citing works with at least one complete pair, one changes its median by the two definitions and five change their five-year recent-reference count. Fourteen of the 25 have at least ten complete pairs. The work-level results do not claim that the median could be estimated for works with no pairs.

Coverage differences matter more than aggregate direction

The four equal-sized rosters do not contribute equal numbers of pairs. Signal Transduction and Targeted Therapy provides 1,789 of the 2,008, or 89.1%. PNAS contributes 603 deposited reference positions and over 500 DOI-resolved target years but one position with both a reference year and a registered year. Its other 524 resolved target years occur where the structured reference year is missing. Consequently the paired table supplies almost no evidence about within-position agreement for that journal. A pooled estimate with the same number of citing works from each journal does not have equal weight by journal when the unit is a reference position. It is important to state this weight explicitly and to keep observed coverage separate from year agreement.

The publication by Yi et al. (2024) supplies a separate exact DOI example. It has 56 positions, including 48 with paired declared and registered years. In two positions the general year is one year earlier and a different typed date field matches the declared year. Neither falls on a consequential boundary for the two defined recency windows. The exact work therefore illustrates a year-field substitution that changes

individual ages without changing the two headline threshold counts. It preserves the distinction between a numeric disagreement and a threshold disagreement.

In that exact example both declarations say 2023, matching the typed print year, whereas both general and online fields say 2022. The resulting two ages, one and two years relative to 2024, remain eligible under both tested windows. In the main sample, all 109 alternative-year matches occur on the print field and none on the online field. Interpreting that pattern as a difference of date roles is more constrained than describing all 124 mismatches as inaccurate references. These observations show which years the public metadata can support. They do not establish which version of the work a researcher read or how its reference field was generated.

Interpretation for age-based summaries

The results support three distinctions in a reproducible age analysis. The first is whether the target can be resolved exactly, given the observed DOI field. The second is whether the two sources provide a usable year for the same position. The third is how sensitive a specific summary, whether a median or a threshold count, is to year substitution on those same positions. Filling 578 missing declared years with registry years would be a separate analytical choice that extends the covered set. An apparent change in the final statistic after such a fill would reflect the larger denominator as well as the source of the year.

For example, a policy retaining every position for which at least one of the two routes gives a year can fill 2,697 of the 2,777 positions. It keeps 111 reference-year positions without a usable exact-target year, 578 positions with only a target year, and 2,008 positions with both. Eighty positions remain unresolved by either route. The number covered by the combined policy is a valid description of available records, but the resulting column necessarily mixes years recovered at different levels in the data model. It may also mix online and print timing. A large increase in apparent analytical coverage under such a policy therefore gives little evidence about whether its resulting age measure is consistent across journals. In the PNAS frame, most of the increase comes from positions without a structured reference year. In the largest of the other frames, the DOI join more often serves to check a year already present.

An explicit print or online year need not be analytically preferable. Each can answer a timing question in its own terms, and even years in a correct-looking format can describe different stages. The exact comparison does not identify which choice better represents intellectual diffusion or the availability of a cited work. It shows instead that provenance decisions can be exposed using the two public records. A consistent citation-clock analysis can keep both the source of the year and the year itself, so that later users can rerun the same threshold using stable inclusion rules. That is especially important where one journal provides many DOI-based years and relatively few years directly in its reference metadata.

Data and code availability

The frozen projection contains the original selected work identities, the DOI and year fields of citation positions and the exact target lookup receipts. Run python scripts/reference_age_1243.py for the comparison summary and python scripts/plot_reference_age_1243.py for the research model and the plot. The saved projection avoids depending on subsequent changes to published date deposits.

References

BARNETT, G. A., FINK, E. L., & DEBUS, M. B. (1989). A Mathematical Model of Academic Citation Age. *Communication Research*, *16*(4), 510-531. https://doi.org/10.1177/009365089016004003

Chow, N. L., Tateishi, N., Goldhar, A., Zaheer, R., Redelmeier, D. A., Cheung, A. H., Schaffer, A., & Sinyor, M. (2023). Does knowledge have a half-life? An observational study analyzing the use of older citations in medical and scientific publications. *BMJ Open*, *13*(5), e072374. https://doi.org/10.1136/bmjopen-2023-072374

Yi, L., Pan, H., Ning, Z., Xu, L., Zhang, H., Peng, L., Liu, Y., Yang, Y., Si, W., Wang, Y., Zhu, X., Huang, S., Meng, Z., & Xie, J. (2024). Clinical and biomarker analyses of SHR-1701 combined with famitinib in patients with previously treated advanced biliary tract cancer or pancreatic ductal adenocarcinoma: a phase II trial. *Signal Transduction and Targeted Therapy*, *9*(1), 347. https://doi.org/10.1038/s41392-024-02052-3

Christen, P. (2012). *Data Matching*. https://doi.org/10.1007/978-3-642-31164-2

Cioffi, A., Coppini, S., Massari, A., Moretti, A., Peroni, S., Santini, C., & Shahidzadeh Asadi, N. (2022). Identifying and correcting invalid citations due to DOI errors in Crossref data. *Scientometrics*, *127*(6), 3593-3612. https://doi.org/10.1007/s11192-022-04367-w

Citation Style Language. (n.d.). *Citation Style Language 1.0.2 Specification*. https://docs.citationstyles.org/en/stable/specification.html

Crossref. (n.d.-a). *Adding references to your metadata record*. https://support.crossref.org/hc/en-us/articles/215578403-Adding-references-to-your-metadata-record

Crossref. (n.d.-b). *References*. https://www.crossref.org/documentation/principles-practices/best-practices/references/

Das, A., & Das, P. (2006). Delay between online and offline issue of journals: A critical analysis. *Library & Information Science Research*, *28*(3), 453-459. https://doi.org/10.1016/j.lisr.2006.03.019

de Jonge, H., & Kramer, B. (2026). Manuscript submission systems and metadata completeness in Crossref: Patterns and associations. *PLOS One*, *21*(3), e0345417. https://doi.org/10.1371/journal.pone.0345417

Delgado-Quirós, L., & Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, *5*(1), 31-49. https://doi.org/10.1162/qss_a_00286

Delgado-Quirós, L., Aguillo, I. F., Martín-Martín, A., López-Cózar, E. D., Orduña-Malea, E., & Ortega, J. L. (2023). Why are these publications missing? Uncovering the reasons behind the exclusion of documents in free-access scholarly databases. *Journal of the Association for Information Science and Technology*, *75*(1), 43-58. https://doi.org/10.1002/asi.24839

Donner, P. (2018). Effect of publication month on citation impact. *Journal of Informetrics*, *12*(1), 330-343. https://doi.org/10.1016/j.joi.2018.01.012

dos Santos, E. A., Peroni, S., & Mucheroni, M. L. (2022). The Way We Cite: Common Metadata Used Across Disciplines for Defining Bibliographic References. *Lecture Notes in Computer Science*, 120-132. https://doi.org/10.1007/978-3-031-16802-4_10

Egghe, L. (2010). A model showing the increase in time of the average and median reference age and the decrease in time of the Price Index. *Scientometrics*, *82*(2), 243-248. https://doi.org/10.1007/s11192-009-0057-3

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, *64*(328), 1183-1210. https://doi.org/10.1080/01621459.1969.10501049

Haustein, S., Bowman, T. D., & Costas, R. (2015). When is an article actually published? An analysis of online availability, publication, and indexation dates. *Proceedings of the 15th International Conference on Scientometrics and Informetrics*. https://www.issi-society.org/proceedings/issi_2015/1170.pdf

Liu, W., Wang, P., & Wang, H. (2026). A matter of time: Publication dates in Scopus. *Journal of Information Science*, *52*(3), 763-771. https://doi.org/10.1177/01655515241293753

Lyu, P., Guo, L., & Zhu, L. (2025). Growth and aging study on literatures of information science & library science based on SSCI data. *Journal of Librarianship and Information Science*, *57*(1), 57-70. https://doi.org/10.1177/09610006231196891

Marx, W., & Bornmann, L. (2016). Change of perspective: bibliometrics from the point of view of cited references—a literature overview on approaches to the evaluation of cited references in bibliometrics. *Scientometrics*, *109*(2), 1397-1415. https://doi.org/10.1007/s11192-016-2111-2

Mesirov, J. P. (2010). Accessible Reproducible Research. *Science*, *327*(5964), 415-416. https://doi.org/10.1126/science.1179653

National Information Standards Organization. (2021). *Dates in Citations*. https://jats.nlm.nih.gov/publishing/tag-library/1.3/chapter/tag-cite-dates.html

Price, D. J. d. S. (1965). Networks of Scientific Papers. *Science*, *149*(3683), 510-515. https://doi.org/10.1126/science.149.3683.510

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology*, *9*(10), e1003285. https://doi.org/10.1371/journal.pcbi.1003285

Stacey, A. G. (2020). Robust parameterisation of ages of references in published research. *Journal of Informetrics*, *14*(3), 101048. https://doi.org/10.1016/j.joi.2020.101048

Tort, A. B. L., Targino, Z. H., & Amaral, O. B. (2012). Rising Publication Delays Inflate Journal Impact Factors. *PLoS ONE*, *7*(12), e53374. https://doi.org/10.1371/journal.pone.0053374

Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, *2*(1), 20-41. https://doi.org/10.1162/qss_a_00112

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., …, Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Zhu, Y. (2025a). An exploratory study on the publication stages of early access articles in different bibliographic databases: A case study of IEEE journals. *PLOS One*, *20*(6), e0325787. https://doi.org/10.1371/journal.pone.0325787

Zhu, Y. (2025b). Evaluation methods and empirical research on comprehensiveness, accuracy and timeliness of early access indexing in the Web of Science. *Journal of Information Science*, 01655515251379429. https://doi.org/10.1177/01655515251379429