Studies in Knowledge and Innovation
Vol. 2 No. 11 (2026)

What the Publication Links in Early Liver Cancer Trial Records Actually Identify

Linda Kennedy · Julian Sullivan · Noah Spencer

Published: November 26, 2026 | Article 1103

Abstract

An early-phase trial can have a publication entered by the registry record holder, a publication linked from PubMed, both, or neither. We compared those routes for 248 ClinicalTrials.gov registrations with completed interventional studies and explicit hepatocellular carcinoma conditions. Ninety-four registrations had at least one PubMed identifier in either route. Of these, 80 were visible only through the automatically derived route and 10 only through entered result references. We then looked for the registration identifier on each linked PubMed record. It appeared in the structured accession field of all 121 derived registration–publication pairs and six of 32 entered pairs. The two routes expose different bibliographic evidence. Identifier agreement reveals how the links can be retrieved and checked, while leaving the contents and relevance of publications for separate assessment.

What the Publication Links in Early Liver Cancer Trial Records Actually Identify

Abstract

An early-phase trial can have a publication entered by the registry record holder, a publication linked from PubMed, both, or neither. We compared those routes for 248 ClinicalTrials.gov registrations with completed interventional studies and explicit hepatocellular carcinoma conditions. Ninety-four registrations had at least one PubMed identifier in either route. Of these, 80 were visible only through the automatically derived route and 10 only through entered result references. We then looked for the registration identifier on each linked PubMed record. It appeared in the structured accession field of all 121 derived registration–publication pairs and six of 32 entered pairs. The two routes expose different bibliographic evidence. Identifier agreement reveals how the links can be retrieved and checked, while leaving the contents and relevance of publications for separate assessment.

The link and the paper are two different observations

A reader searching the publications section of a trial registry might see several titles reported under one trial number. Another reader starting from a literature database may see the same trial number in a citation field or in an abstract. Both routes are useful. They also carry different information about who provided the link and what can be deduced from it. An entered reference can display an article that does not mention the record's identifier in PubMed. An automatically derived reference can display a paper that mentions the identifier for a reason other than providing the study's main report. Treating all visible links as endorsements of the same kind can obscure these distinctions before anyone opens an article.

The question matters because registration and publication serve related but separate transparency goals. Simes (1986) explained how trial registration can address the selective visibility of research. De Angelis et al. (2004) articulate an editorial rationale for registering trials, and the World Health Organization describes a core dataset for registry records (World Health Organization, n.d.). Ross et al. (2009) and Ross et al. (2012) examined publication following registration in different trial cohorts. Yerunkar et al. (2025) situate links, results, and public-facing information within a broader assessment of transparency. These works motivate locating a study across public records. In the present analysis the measurable outcome is the presence of particular structured links, rather than the availability of every published report about a trial.

An indexed pilot paper on refractory advanced hepatocellular carcinoma by Yi et al. (2024) illustrates a search beginning from a known report. Its PubMed identifier, 38325225, uniquely identifies that journal article. Our saved PubMed metadata for it contains an abstract and contains no ClinicalTrials.gov number in either of the two fields used in this study. The paper is consequently useful as a concrete example of why identifying a publication by its bibliographic key and finding it by a registration key are separate tasks. Its identifier did not appear in this analysis's extracted link fields. We treat it as a worked example outside the cohort counts and make no claim about which registries may describe its underlying trial. Analysis of the intervention or the reported study outcomes is outside the design of this paper.

Previous studies of link ascertainment

Huser and Cimino (2012) examine the performance of links in opposite directions between ClinicalTrials.gov and PubMed, and Huser and Cimino (2013) use both sources to study the dissemination of trial information. Bashir et al. (2017) distinguish automated from manual steps used to link a trial with literature about it. Salholz-Hillel et al. (2022) show that source-specific differences in indexing and entered references can matter for a cohort of published trials. Liu et al. (2022) and Goodwin et al. (2018) go a step further by proposing ways to find likely matches that do not already share a visible identifier. These are identification and retrieval problems. Our design instead marks the route and tests the presence of a common identifier for the subset of pairs that the registry itself publishes. It can be reproduced without judging article topics or inferring unreported matches.

The registry distinguishes general references from references marked as relating to study results by the person making the entry (National Library of Medicine, n.d.-b). It also displays literature found through matching in PubMed and specifies that this category may include publications that are not about the study itself (National Library of Medicine, n.d.-a). The three categories create distinct source relationships. We exclude general references from the main two-route comparison because they need not purport to convey findings. A report credited to a study can still require detailed inspection before it is suitable for evidence synthesis. Hunter et al. (2022) explain why searches in registries should be planned and documented in systematic reviews. The retrieval output in this audit could be the start of such a search, but the analysis does not replace content assessment or a search of other registries or databases.

Sampling the registrations

We queried the public ClinicalTrials.gov API for interventional studies returned under hepatocellular carcinoma, marked complete, completed from 2014 through 2024, and classified in phase I or II, including early phase I. The API snapshot yielded 352 unique registration IDs. To exclude records returned indirectly by the search, we retained only records whose declared condition names contained hepatocellular carcinoma, hepatocellular cancer, or the standalone abbreviation HCC. This removed 101 records. We used the observed completion date type and retained studies with an actual rather than estimated completion in the date window. After applying the phase rules we removed three records outside the allowed standalone or combination designations. The final dataset comprised 248 registrations: 80 phase I or early phase I, 38 combined phase I/II, and 130 phase II.

The filters describe the selected registry records. The diagnosed condition of any individual is not recorded in the extract. A structured phase I/II design is kept separate from a standalone phase II design rather than assigned to both. Published registry records can be updated and may have multiple records in different systems. Tse et al. (2018) discuss how interpretation depends on the definition and timing of fields. Miron et al. (2020) give additional examples of obstacles to reusing the data as organized for public viewing. We check uniqueness of the public NCT identifier and freeze one retrieval. This is a less ambitious linkage than deduplicating a registry ecosystem, as discussed by Thiele et al. (2021), or reconciling conflicting information across records for one trial, as in Speich et al. (2021).

Two units of comparison

At the registration level, we assign each NCT record to a cross-classification of two yes-or-no fields. One flag requires a positive integer PubMed identifier in a reference labeled RESULT. We refer to this as the entered-result route. The other flag requires a valid identifier labeled DERIVED. Every selected trial has exactly one of four states, with both routes and neither allowed. Multiple PMIDs in one route increase the number of pairs but leave the registration flag at one. We also compare the sets of identifiers across the two routes for the same NCT record, because coexistence of the routes does not guarantee that both point to a common paper.

At the pair level, we deduplicate each combination of registration number, PubMed identifier, and route and check the current public PubMed metadata. The request used standard E-utilities and retrieved XML for 149 unique PubMed identifiers linked in the cohort and for the separate example publication. A matching accession is the exact NCT token in PubMed's DataBank accession fields. We separately check whether the same token appears in the title or the abstract. The title and abstract are combined for this test so missing abstracts remain eligible for a title match. We save indicators and identifier lists, leaving title and abstract text outside the archived extract. The result measures exact agreement between fields. This approach, like the grouping distinctions in construction of AACT (Tasneem et al., 2012), depends on retaining the relation between each registration and its own linked reference.

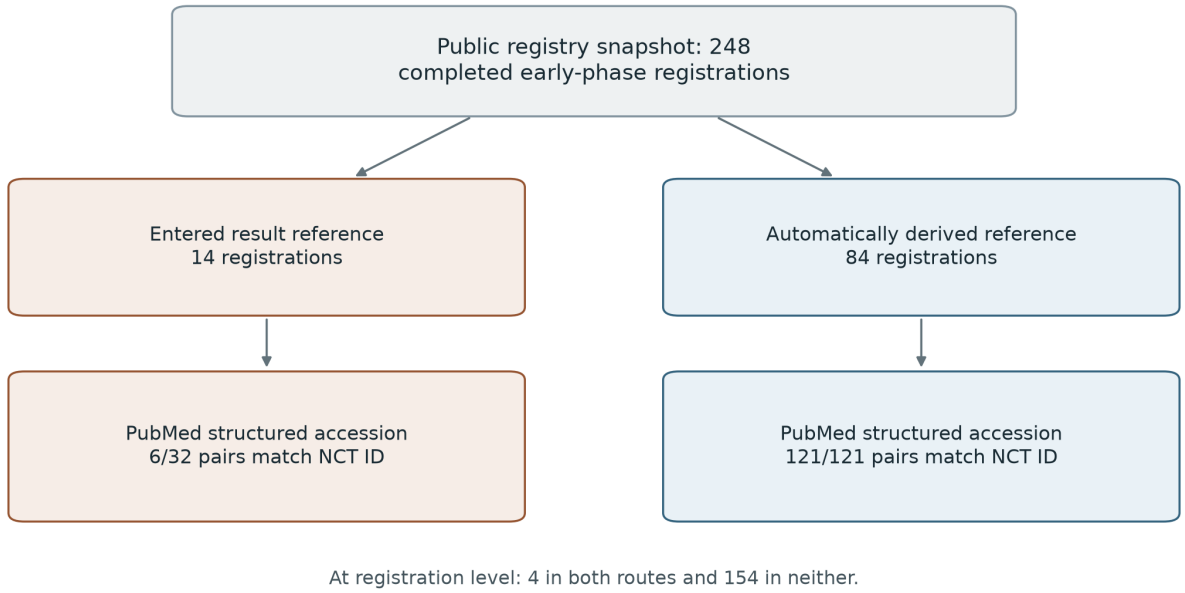


Figure 1. The first split describes visibility by route within the selected cohort. The second stage checks the exact registered identifier against structured PubMed accession data for each unique pair recorded in that route.

What the two routes contribute

Ninety-four of 248 registrations had at least one identifier in either route, or 37.9%. Eighty were visible only in the derived route. Ten were visible only in the entered-result route, and four had both. Using the derived route alone would retain 84 registrations, compared with 14 for the entered-result route alone. Counting both therefore adds 10 registrations beyond the larger set, a gain of 4.0 percentage points of the 248 registrations. The analysis counts 154 records as having neither link type. Of these, 16 had at least one background-type reference. The distinction explains why total links shown in a browser can exceed the two route flags, even when the same studies are inspected.

Table 1. Cross-classification of registration-level link availability. Phase and completion period are two alternative partitions of the same 248 records.

Phase or actual completion window	Registrations	Entered result only	Derived only	Both	Neither
All included registrations	248	10	80	4	154
Early phase I or phase I	80	4	18	0	58
Combined phase I/II	38	1	16	2	19
Standalone phase II	130	5	46	2	77
Completed in 2014–2018	89	6	25	1	57
Completed in 2019–2024	159	4	55	3	97

All four records displaying both routes had disjoint PubMed identifier sets between them. That observation is limited to category entries exposed by the API at retrieval. It does not show how the registry handles overlap during ingestion or display. If duplicate links are resolved upstream, an audit of the displayed records will reflect what survived that processing. A study tracing possible mechanisms would need dated public exports or authoritative source documentation. The phase and calendar partitions illustrate how the counts can be summarized, but we do not interpret the difference across cohorts as a temporal change in publication or a phase-related effect. The evidence for any individual paper still depends on the publication and the reason for the link.

Table 2. Comparisons of the exact NCT token in retrieved PubMed metadata. The pair counts differ from registration counts because one registration may contain more than one article identifier.

PubMed metadata check	Entered-result pairs	Automatically derived pairs
Unique registration–PMID pairs	32	121
Same NCT token in structured accession	6	121
Same token in title or abstract	5	84
Same token in both field types	5	84
Other NCT token only	1	0

All 150 PubMed records requested were retrieved. The separate example is excluded from the pair table. Every derived pair contains the expected NCT token in a structured accession. In 37 of the 121 pairs, that token is absent from the visible title and abstract text returned by PubMed. Requiring text alone would therefore lose those exact matches despite the structured linkage. Among entered-result pairs, six have a matching accession and five also have the token in text. Twenty-six lack the expected token in both fields. In one of these the recorded PubMed metadata has a different NCT number. We do not infer that this was a miscitation. A paper may describe more than one study, a registry record may link a secondary analysis, and the exact identifier may be available elsewhere. The title and abstract test provides a reproducible measure of metadata visibility without deciding those cases.

Five derived pairs point to the same indexed paper, PMID 36852452, from five distinct registrations. In each pair the corresponding registration number appears in that paper's structured PubMed metadata. The other linked PMIDs occur once each in the cohort's pair inventory. This gives 149 distinct PMIDs across 153 pairs. The duplicate at the publication level is therefore compatible with distinct valid identifier matches at the pair level. An archive organized as one row per paper would have to retain all five

associations in a linked table. Dropping extra rows after sorting by the PMID would reduce the count of observed registration-to-publication relationships without resolving what each relationship meant.

What a future literature search should retain

The main contrast shows that reliance on one route changes the set of registrations for which a paper is immediately discoverable. The accession check helps explain why a discovery tool that searches only article text sees fewer exact matches than one using structured identifiers. It cannot distinguish the main report, a companion paper, or a reference to a different trial. The absence of a visible identifier also cannot establish whether any results were published. Riveros et al. (2013) compare the information available across registries and journals. JATS4R (2018) specifies separate metadata for the registry and the registered identifier in article XML. Chan et al. (2004) and Mathieu (2009) show why later evaluations of published outcomes require inspection of documents and definitions. An identifier audit can supply a transparent queue for that inspection, without substituting for it.

The analysis also cannot determine compliance with rules for posting results. Goldacre et al. (2018) assess a defined obligation using a specific registry and eligible cohort. In the current data there is only a public snapshot of visible reference fields, with no reconstruction of each reporting obligation or its dates. Similarly, the fractions shown here cannot stand in for sensitivity and specificity of publication matching, which would require a reference standard for all 248 registrations. To support a follow-up, the archived pair file preserves the provenance of the entered and automatically derived identifiers and the exact-field indicators. Search documentation standards such as (Rethlefsen et al., 2021) and reporting guidance for observational studies using routine data (Benchimol et al., 2015) motivate keeping the denominators and decision rules visible.

Each service can change independently. An added bibliographic accession could make a record discoverable in PubMed while leaving an older electronic export from the registry unchanged. Conversely, a newly entered reference might be presented before the bibliographic record has comparable indexing. The retrieval receipts identify the versions that were combined here. The design therefore yields a current cross-section of the two services. It leaves the lag between publication, accession assignment, and public display to an analysis with archived versions. That distinction also matters when comparing the earlier and later completion groups, since a stopping date governs what could be found for each group.

Registration-level reach and link-level attribution support different follow-up work. A reader interested in whether a registered study is traceable needs to check which registrations have any candidate. A reviewer assessing the evidence for a particular candidate needs the exact registration and PMID pair, including the route that supplied it. Neither counting method requires reading or inferring a treatment result. Retaining the route makes it possible to review a manually assigned result identifier with the appropriate record context while checking the more numerous automatic associations against their own accession fields.

Data and code availability

The registry snapshot and PubMed identifier extract are sufficient to regenerate both tables and the model figure. The registry analysis produces the registration-level flags and underlying link sets. The PubMed metadata analysis uses the paired identifiers and produces the exact-match indicators. The figure script reads those outputs. Source requests and hashes separate retrieved evidence from derived indicators, in keeping with general guidance on reusable research data (Wilkinson et al., 2016) and on recording workflow provenance (Leo et al., 2024). The study evidence comes from publicly issued records and contains no individual clinical measurements.

References

Bashir, R., Bourgeois, F. T., & Dunn, A. G. (2017). A systematic review of the processes used to link clinical trial registrations to their published results. *Systematic Reviews*, *6*(1), 123. https://doi.org/10.1186/s13643-017-0518-3

Benchimol, E. I., Smeeth, L., Guttmann, A., Harron, K., Moher, D., Petersen, I., Sørensen, H. T., von Elm, E., Langan, S. M., & RECORD Working Committee (2015). The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Medicine*, *12*(10), e1001885. https://doi.org/10.1371/journal.pmed.1001885

Chan, A., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials. *JAMA*, *291*(20), 2457. https://doi.org/10.1001/jama.291.20.2457

De Angelis, C., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., Kotzin, S., Laine, C., Marusic, A., Overbeke, A. J. P., Schroeder, T. V., Sox, H. C., & Weyden, M. B. V. D. (2004). Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine*, *351*(12), 1250-1251. https://doi.org/10.1056/nejme048225

Goldacre, B., DeVito, N. J., Heneghan, C., Irving, F., Bacon, S., Fleminger, J., & Curtis, H. (2018). Compliance with requirement to report results on the EU Clinical Trials Register: cohort study and web resource. *BMJ*, k3218. https://doi.org/10.1136/bmj.k3218

Goodwin, T. R., Skinner, M. A., & Harabagiu, S. M. (2018). Automatically Linking Registered Clinical Trials to their Published Results with Deep Highway Networks. *AMIA Joint Summits on Translational Science Proceedings*, *2018*, 54–63. https://pubmed.ncbi.nlm.nih.gov/29888040/

JATS4R (2018). *Clinical trials*. https://jats4r.niso.org/clinical-trials/

Hunter, K. E., Webster, A. C., Page, M. J., Willson, M., McDonald, S., Berber, S., Skeers, P., Tan-Koay, A. G., Parkhill, A., & Seidler, A. L. (2022). Searching clinical trials registers: guide for systematic reviewers. *BMJ*, *377*, e068791. https://doi.org/10.1136/bmj-2021-068791

Huser, V., & Cimino, J. J. (2012). Precision and Negative Predictive Value of Links between ClinicalTrials.gov and PubMed. *AMIA Annual Symposium Proceedings*, *2012*, 400–408. https://pubmed.ncbi.nlm.nih.gov/23304310/

Huser, V., & Cimino, J. J. (2013). Linking ClinicalTrials.gov and PubMed to Track Results of Interventional Human Clinical Trials. *PLoS ONE*, *8*(7), e68409. https://doi.org/10.1371/journal.pone.0068409

Leo, S., Crusoe, M. R., Rodríguez-Navas, L., Sirvent, R., Kanitz, A., De Geest, P., Wittner, R., Pireddu, L., Garijo, D., Fernández, J. M., Colonnelli, I., Gallo, M., Ohta, T., Suetake, H., Capella-Gutierrez, S., de Wit, R., Kinoshita, B. P., & Soiland-Reyes, S. (2024). Recording provenance of workflow runs with RO-Crate. *PLOS ONE*, *19*(9), e0309210. https://doi.org/10.1371/journal.pone.0309210

Liu, S., Bourgeois, F. T., & Dunn, A. G. (2022). Identifying unreported links between ClinicalTrials.gov trial registrations and their published results. *Research Synthesis Methods*, *13*(3), 342-352. https://doi.org/10.1002/jrsm.1545

Mathieu, S. (2009). Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials. *JAMA*, *302*(9), 977. https://doi.org/10.1001/jama.2009.1242

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. https://doi.org/10.1016/j.esmoop.2024.102239

Miron, L., Gonçalves, R. S., & Musen, M. A. (2020). Obstacles to the reuse of study metadata in ClinicalTrials.gov. *Scientific Data*, *7*(1), 443. https://doi.org/10.1038/s41597-020-00780-z

National Library of Medicine. (n.d.-a). *How to Read Study Results*. https://clinicaltrials.gov/study-basics/how-to-read-study-results

National Library of Medicine. (n.d.-b). *Protocol Registration Data Element Definitions for Interventional and Observational Studies*. https://clinicaltrials.gov/policy/protocol-definitions

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., PRISMA-S Group, Blunt, H., Brigham, T., Chang, S., Clark, J., Conway, A., Couban, R., de Kock, S., Farrah, K., Fehrmann, P., Foster, M., Fowler, S. A., …, Young, S. (2021). PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Systematic Reviews*, *10*(1), 39. https://doi.org/10.1186/s13643-020-01542-z

Riveros, C., Dechartres, A., Perrodeau, E., Haneef, R., Boutron, I., & Ravaud, P. (2013). Timing and Completeness of Trial Results Posted at ClinicalTrials.gov and Published in Journals. *PLoS Medicine*, *10*(12), e1001566. https://doi.org/10.1371/journal.pmed.1001566

Ross, J. S., Mulvey, G. K., Hines, E. M., Nissen, S. E., & Krumholz, H. M. (2009). Trial Publication after Registration in ClinicalTrials.Gov: A Cross-Sectional Analysis. *PLoS Medicine*, *6*(9), e1000144. https://doi.org/10.1371/journal.pmed.1000144

Ross, J. S., Tse, T., Zarin, D. A., Xu, H., Zhou, L., & Krumholz, H. M. (2012). Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis. *BMJ*, *344*(jan03 1), d7292-d7292. https://doi.org/10.1136/bmj.d7292

Salholz-Hillel, M., Strech, D., & Carlisle, B. G. (2022). Results publications are inadequately linked to trial registrations: An automated pipeline and evaluation of German university medical centers. *Clinical Trials*, *19*(3), 337-346. https://doi.org/10.1177/17407745221087456

Simes, R. J. (1986). Publication bias: the case for an international registry of clinical trials. *Journal of Clinical Oncology*, *4*(10), 1529-1541. https://doi.org/10.1200/jco.1986.4.10.1529

Speich, B., Gloy, V. L., Klatte, K., Gryaznov, D., Taji Heravi, A., Ghosh, N., Marian, I. R., Lee, H., Mansouri, A., Lohner, S., Saccilotto, R., Nury, E., Chan, A., Blümle, A., Odutayo, A., Hopewell, S., & Briel, M. (2021). Reliability of Trial Information Across Registries for Trials With Multiple Registrations. *JAMA Network Open*, *4*(11), e2128898. https://doi.org/10.1001/jamanetworkopen.2021.28898

Tasneem, A., Aberle, L., Ananth, H., Chakraborty, S., Chiswell, K., McCourt, B. J., & Pietrobon, R. (2012). The Database for Aggregate Analysis of ClinicalTrials.gov (AACT) and Subsequent Regrouping by Clinical Specialty. *PLoS ONE*, *7*(3), e33677. https://doi.org/10.1371/journal.pone.0033677

Thiele, C., Hirschfeld, G., & von Brachel, R. (2021). Clinical trial registries as Scientometric data: A novel solution for linking and deduplicating clinical trials from multiple registries. *Scientometrics*, *126*(12), 9733-9750. https://doi.org/10.1007/s11192-021-04111-w

Tse, T., Fain, K. M., & Zarin, D. A. (2018). How to avoid common problems when using ClinicalTrials.gov in research: 10 issues to consider. *BMJ*, k1452. https://doi.org/10.1136/bmj.k1452

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

World Health Organization. (n.d.). *Trial Registration Data Set*. https://www.who.int/clinical-trials-registry-platform/network/who-data-set

Yerunkar, S. S., Hildebrand, N., & Strech, D. (2025). Towards monitoring of global health research: an exploratory analysis of transparency and stakeholder engagement. *BMJ Open*, *15*(12), e102010. https://doi.org/10.1136/bmjopen-2025-102010