Uncertainty and Assessment
Volume 1 · Number 1 · 2026

Matching Trial Names Is Not Matching Populations: Enrollment Reconciliation in an HCC Evidence Table

Christopher Webb[1] · Gabriel Wells[2] · Jason Reed[3]

[1] New Zealand · [2] Switzerland · [3] Sweden

Original publication: October 9, 2026 | Article 576

Abstract

Seven of 22 linked enrollment totals matched exactly between an HCC review table and ClinicalTrials.gov snapshots. Nevertheless, their log-count regression had an R-squared of 0.987. Four relative discrepancies exceeded ten percent. Case reconciliation showed why disagreement cannot automatically be called error: ORIENT-32 combines a safety run-in with a randomized phase, and HIMALAYA's registry includes an additional treatment arm. Two other review rows lacked registry identifiers and remained outside the paired analysis. This audit favors explicit phase, arm, and population labels over a single numerical agreement check when reusing trial counts.

Study-level reporting-audit draft using a published review and public registry snapshots. No patient-level data, treatment-effect estimate, or clinical recommendation is produced.

Matching Trial Names Is Not Matching Populations: Enrollment Reconciliation in an HCC Evidence Table

Study-level reporting-audit draft using a published review and public registry snapshots. No patient-level data, treatment-effect estimate, or clinical recommendation is produced.

Abstract

Seven of 22 linked enrollment totals matched exactly between an HCC review table and ClinicalTrials.gov snapshots. Nevertheless, their log-count regression had an R-squared of 0.987. Four relative discrepancies exceeded ten percent. Case reconciliation showed why disagreement cannot automatically be called error: ORIENT-32 combines a safety run-in with a randomized phase, and HIMALAYA's registry includes an additional treatment arm. Two other review rows lacked registry identifiers and remained outside the paired analysis. This audit favors explicit phase, arm, and population labels over a single numerical agreement check when reusing trial counts.

1. The denominator behind a study identifier

A registry identifier can connect a published trial report with a public study record. That connection is useful, but it does not guarantee that every numerical field has the same denominator. A registry may describe the whole study, whereas a review selects randomized comparisons, a particular phase, or a subset of treatment arms. Counting a successful identifier match as a successful sample-size verification can therefore produce a misleading audit, even when both linked records faithfully describe their intended populations.

The evidence table examined here accompanies a network meta-analysis of first-line immunotherapy and targeted therapy in advanced hepatocellular carcinoma, including subgroup analyses by infection status (Li et al., 2026). We use its study-characteristics table as an observable reporting artifact. The research question is whether review-reported enrollment totals agree numerically with linked registry snapshots, and what a discrepancy can identify once phase and arm boundaries are considered. This is an audit of records rather than a new synthesis of the clinical effects discussed in the source review.

Two propositions organize the calculations. P1 anticipates that larger registry studies tend to have larger totals in the review table. P2 describes the mean log ratio of review counts to registry totals. Neither proposition states that a ratio below one is an error or a marker of biased treatment evidence. We combine the calculations with two case reconciliations because a numerically small discrepancy and a conceptually wrong denominator are different problems. A high correlation can coexist with either.

2. Why a reporting audit needs several sources

Systematic-review reporting guidance and its network-meta-analysis extension emphasize identifying studies and specifying the evidence being synthesized (Page et al., 2021; Hutton et al., 2015). That task includes connecting multiple reports to their underlying trial. It also requires preserving the distinction between a trial and an analysis population. In a multi-arm network, including a subset of comparisons can create a denominator that is smaller than the total registered enrollment without introducing an extraction mistake.

Network methods address dependence and consistency among treatment comparisons, while heterogeneity measures describe variation in effect estimates (Higgins et al., 2012; Higgins and Thompson, 2002). Neither provides a substitute for checking which participants contributed to a quoted

count. Our audit therefore stays upstream of those models. We examine record-level counts, arm lists, and enrollment-type fields. We do not translate the resulting differences into a revised heterogeneity statistic, a treatment ranking, or a judgment that a clinical network is inconsistent.

The ClinicalTrials.gov results-database literature explains the role of structured public study information (Zarin et al., 2011). Research on selective outcome reporting also demonstrates why published and prespecified study information deserve comparison (Chan et al., 2004). Those concerns motivate careful reconciliation, but they do not make every numerical mismatch evidence of selective reporting. Our two-source linkage can identify a question for review. Establishing its reason requires examining the scope of each field and, where available, the underlying trial publication.

Evidence-grading frameworks distinguish certainty of evidence from the presence of a reporting detail (Guyatt et al., 2008; Guyatt et al., 2011). We do not assign a certainty grade based on whether enrollment counts match. A review count can agree perfectly with a registry and still refer to the wrong outcome population; it can disagree for a legitimate reason documented in the trial report. This distinction is important when an automated record check is used to prioritize human verification rather than replace scientific assessment.

3. A fixed review table and a dated registry snapshot

The source table has 24 study rows. We preserve the printed sample-size cells and sum the slash-separated arm counts within each row. Twenty-two rows provide an identifier matching the NCT format; two state that the identifier is not reported. We download the 22 corresponding public study records through the ClinicalTrials.gov API and preserve each response with its retrieval timestamp and checksum (National Library of Medicine, 2026b). All linked identifiers match the requested record identifiers. The acquisition date is November 6, 2026.

The public API exposes enrollment counts, enrollment types, arm-group listings, and record-update fields in structured modules (National Library of Medicine, 2026a). We extract the design-module enrollment total and distinguish ACTUAL from ESTIMATED. Twenty records use the former label and two the latter. These are the latest captured records, not reconstructed snapshots from the review's search date. Their last submitted updates vary across trials, so a common download date does not imply a common date on which trial information was last revised.

Rows without an identifier remain in the extraction ledger but outside the paired count analysis. We do not infer an identifier from a similar title or enter zero as a missing registry count. Missing-data methodology provides a reason to preserve this distinction (Little and Rubin, 2002). The paired sample consequently comprises 22 linked records rather than all 24 table rows. Its conclusions are conditional on this documented linkage rule and do not establish how readily every omitted identifier could be recovered by an expert search.

We also retain the number of sample-size components, the number of printed randomization-ratio components, and the number of registry arm groups. These fields need not agree in every study, because phase, cohort, and comparator definitions can differ. Recording them makes the discrepancy inspectable. Datasheet principles encourage preserving provenance and variable definitions (Gebru et al., 2021), while FAIR principles concern making research information findable and reusable under clear conditions (Wilkinson et al., 2016). Neither principle permits replacing an unresolved value with a plausible correction.

4. Two numerical summaries and a scope check

For each linked record, the first model regresses the natural logarithm of the summed review count on the natural logarithm of registry enrollment. It includes an intercept and uses HC3 standard errors. Heteroskedasticity-consistent inference concerns the model's residual variance (White, 1980); it does not establish that the 22 selected trials are a random sample of all possible cancer studies. P1 is a descriptive scale relationship within this fixed evidence table. A slope near one is compatible with several kinds of systematic or case-specific disagreement.

For P2, we calculate the natural logarithm of the review-to-registry count ratio for each record. An intercept-only model estimates its arithmetic mean on the log scale, with a conventional Student interval over the 22 records. We repeat this calculation among the 20 ACTUAL-enrollment records, retaining the distinction between realized and planned counts. These are study-weighted summaries, not patient-weighted estimates. A large trial and a small trial each contribute one log ratio to the reported mean.

Three additional descriptive quantities make the comparison less abstract: the number of exact enrollment matches, the median absolute difference divided by registry enrollment, and the number exceeding a ten-percent relative difference. Ten percent is a reporting threshold for flagging cases, not a validated clinical or data-quality cutoff. A concordance framework explicitly addresses agreement rather than ordinary correlation (Lin, 1989). We do not calculate a concordance coefficient here; instead we show why a high regression fit leaves the actual population definitions unresolved.

The record-level design appears in Figure 1. Counts feed the numerical summaries, while phase and arm information qualify their interpretation. The scope check is not an additional regressor with an estimated causal effect. Causal-diagram methodology distinguishes an assumed relationship from identification of an intervention (Greenland et al., 1999). Similarly, matching methods require a defined comparison population beyond a shared label (Stuart, 2010). Our identifier linkage matches study records, not patients, randomized groups, or exchangeable clinical populations.

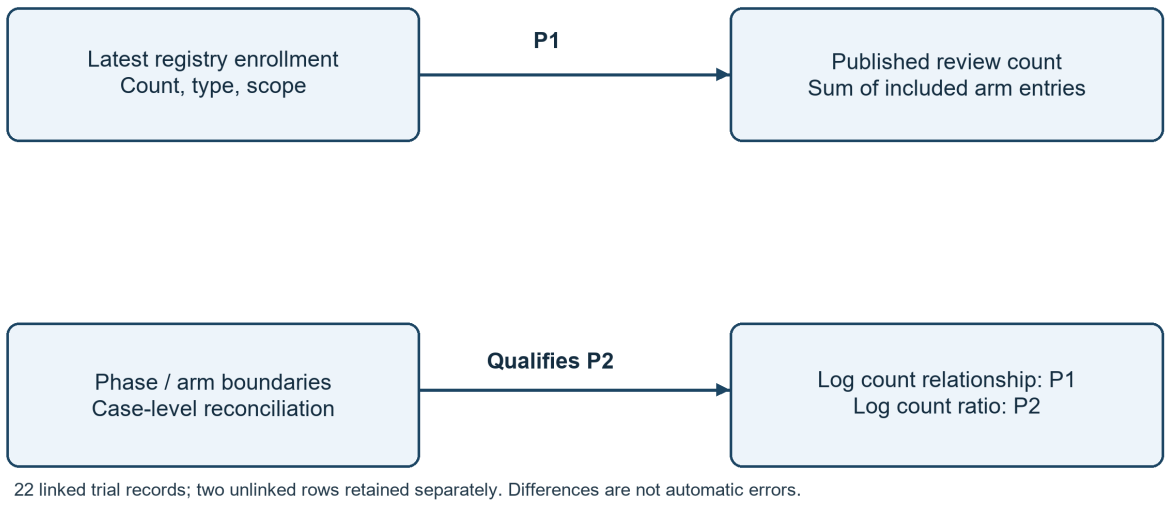

Figure 1. P1 examines the log-count relationship; P2 summarizes log count ratios. Scope reconciliation explains what a discrepancy may mean. The observational unit is a linked trial record.

5. Numerical agreement is incomplete despite high fit

Seven of the 22 enrollment totals match exactly. The median absolute relative difference is 2.77%, and four records exceed the ten-percent reporting threshold. The number of registry arm groups differs from the number of review sample-size components in three linked records. These calculations identify where the two structured representations differ. They are not a count of confirmed review errors. The largest relative enrollment difference occurs in BRISK-FL, for which the review sum is 1,155 and the captured registry total is 1,714; its explanation remains unresolved in this audit.

The log-count regression has a slope of 0.9966, with a standard error of 0.0391 and a 95% interval from 0.9150 to 1.0781. Its intercept is -0.0339, and R-squared is 0.9867. Thus trial size is closely associated across sources. This is expected to some extent because the records describe the same studies and span different sizes. It does not show that the two count fields refer to the same enrolled, randomized, treated, or analyzed participants within each trial.

Table 1. Record-level estimates. The log-count regression uses an intercept and HC3 covariance. The other models estimate mean log count ratios with conventional Student uncertainty across linked records; no clinical outcome or treatment-effect model is fitted. Intercept-only R-squared is zero apart from numerical rounding.

Model	Term	Estimate (SE)	95% CI	N
Log review enrollment versus log registry total	Log registry enrollment: P1	0.9966 (0.0391)	[0.9150, 1.0781]	22
Log review enrollment versus log registry total: fit	R-squared = 0.9867	-	-	22
Mean log review-to-registry ratio	Mean log enrollment ratio: P2	-0.0549 (0.0223)	[-0.1013, -0.0086]	22
Mean log review-to-registry ratio: fit	R-squared = 0.0000	-	-	22
Actual-enrollment records only	Mean log enrollment ratio: P2	-0.0643 (0.0234)	[-0.1132, -0.0154]	20
Actual-enrollment records only: fit	R-squared = 0.0000	-	-	20

The mean log ratio is -0.0549, with a standard error of 0.0223 and a 95% interval of [-0.1013, -0.0086]. Restricting the calculation to ACTUAL-enrollment records gives -0.0643, with an interval of [-0.1132, -0.0154]. These negative averages describe smaller review totals on the chosen scale. They do not estimate missing-patient bias or prove systematic under-extraction. Removing estimated records addresses one field-type distinction while leaving phase, arm, and analysis-population differences intact.

The intervals assume a hypothetical sampling interpretation across comparable trial records. Related trials can share sponsors, investigators, or reporting practices, and the current covariance does not model all such dependence. Cluster-robust guidance makes the grouping decision substantive rather than automatic (Cameron and Miller, 2015). We do not invent a sufficient collection of independent sponsor clusters from the small table. The observed finite-table counts remain directly reproducible even if their generalization to other evidence collections is uncertain.

6. Two cases where the denominator can be reconciled

ORIENT-32 supplies a clear phase example. Its primary report states that 595 participants were enrolled: 24 in a safety run-in and 571 randomized in the later phase, with 380 and 191 assigned to the comparison groups (Ren et al., 2021). The registry total of 595 and the review sum of 571 therefore describe different, documented boundaries. Replacing the review's randomized comparison count with the full registry enrollment would insert the run-in participants into a denominator that did not originally contain them.

HIMALAYA illustrates an arm boundary. The captured registry lists four arms and 1,324 randomized participants; its participant-flow record includes an additional 153-participant regimen group. The primary report describes 1,171 participants in the three reported regimens, with counts of 393, 389, and 389 (Abou-Alfa et al., 2022). That total matches the review sum. The 153-person difference is therefore reconcilable at the arm level. A strict automated equality rule would flag the row without recognizing why its selected comparison population is smaller.

These examples do not resolve every discrepancy. BRISK-FL, the smaller nintedanib study, and other nonmatching rows still require publication-specific or registry-history checks. We retain their original counts rather than generalizing the explanation from the two resolved cases. Measurement-error analysis cautions against assigning a single interpretation to differences between operational measurements (Carroll et al., 2006). Here the operational distinction is particularly concrete: enrollment fields and selected-arm counts can name different sets of people while carrying the same trial identifier.

An internal structural flag is separate from those legitimate population boundaries. The ORIENT-32 sample cell contains two arm counts, but its printed ratio contains three components. The primary report describes two-to-one randomization. That inconsistency calls for reconciliation with the publication; it is not fixed silently in the extracted table. Keeping an internal formatting or transcription concern separate from an explained external count difference makes the audit more useful than a single binary agreement label.

7. Interpretation without clinical overreach

Study-level patterns cannot be transferred directly to individual patients. The ecological-correlation problem provides a longstanding warning about moving between levels of analysis (Robinson, 1950). Our mean log ratio is a property of records, not an estimate of any patient's likelihood of exclusion or benefit. A missing participant identifier, event count, or outcome population cannot be supplied by this regression. The audit therefore has no basis for recommending one therapy over another or modifying a clinical effect estimate.

Statistical significance also does not classify the reason for disagreement. The ASA statement on p-values emphasizes interpretation beyond a threshold, and work on false research findings highlights the consequences of design and analytic choices (Wasserstein and Lazar, 2016; Ioannidis, 2005). A negative mean with an interval below zero remains compatible with legitimate subset selection. Conversely, a mean near zero could hide offsetting errors. Case evidence and field definitions are necessary to interpret the arithmetic rather than assigning quality from the sign of a test.

Checking many fields could create a family of potential findings. False-discovery-rate methods and their extension under dependence address multiplicity when such a family is defined (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). We report both count summaries and the retained structural flags, with no search for a favorable clinical endpoint. Influence diagnostics could show whether a large discrepancy dominates a fitted relationship (Cook, 1977), but deleting that case without resolving its scope would weaken the audit. The unresolved rows remain visible in the delivered ledger.

8. A reusable reconciliation record

An improved extraction form would retain an enrollment number together with its role: planned, enrolled, randomized, treated, or analyzed; phase and selected arms would be explicit. The registry version and publication locator should travel with that role. Reusing an identifier without these fields leaves later reviewers to repeat the denominator search. This proposed record structure follows from the observed ambiguity; it is not claimed to be a prospectively validated tool for detecting review errors.

Future validation should reserve a separately chosen evidence collection and predefine how discrepancies will be adjudicated. Model-selection bias and structured-validation guidance explain why repeatedly adjusting rules on the same small table can overstate their subsequent accuracy (Varma and Simon, 2006; Roberts et al., 2017). The current work supplies calculations and two resolved examples, while leaving broader detection performance unestimated. Human scientific review remains necessary for the clinical relevance of a selected population and the interpretation of any unresolved record conflict.

9. Conclusion

The two sources track study size closely but coincide exactly on only seven linked totals. Phase and arm reconciliation explains why at least some nonmatches are appropriate, while an internal ratio-format inconsistency deserves a different kind of follow-up. A useful denominator audit therefore records both the number and the population it names. High correlation is insufficient, and a discrepancy is a starting point for verification rather than a clinical judgment or an automatic declaration of error.

References

1. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
2. National Library of Medicine. (2026a). ClinicalTrials.gov API: Study data structure and data field definitions. https://clinicaltrials.gov/data-api/api.
3. Benjamini, Yoav and Daniel Yekutieli. “The control of the false discovery rate in multiple testing under dependency.” *The Annals of Statistics*, vol. 29, no. 4, 2001, pp. 1165-1188. https://doi.org/10.1214/aos/1013699998.
4. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
5. Higgins, Julian P. T.; Thompson, Simon G.. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in Medicine. https://doi.org/10.1002/sim.1186.
6. Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.
7. Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1-21. https://doi.org/10.1214/09-sts313.
8. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
9. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.
10. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.
11. Guyatt, Gordon H, et al. "GRADE: an emerging consensus on rating quality of evidence and strength of recommendations." *BMJ*, vol. 336, no. 7650, 2008, pp. 924-926. https://doi.org/10.1136/bmj.39489.470347.ad.
12. National Library of Medicine. (2026b). ClinicalTrials.gov study record snapshots for 22 hepatocellular-carcinoma trial identifiers, retrieved November 6, 2026. https://clinicaltrials.gov/.
13. Abou-Alfa, Ghassan K; Lau, George; Kudo, Masatoshi; Chan, Stephen L; Kelley, Robin Kate; Furuse, Junji; Sukeepaisarnjaroen, Wattana; Kang, Yoon-Koo; Van Dao, Tu; De Toni, Enrico N; Rimassa, Lorenza; Breder, Valeriy; Vasilyev, Alexander; Heurgué, Alexandra; Tam, Vincent C; Mody, Kabir; Thungappa, Satheesh Chiradoni; Ostapenko, Yuriy; Yau, Thomas; Azevedo, Sergio; Varela, María; Cheng, Ann-Lii; Qin, Shukui; Galle, Peter R; Ali, Sajid; Marcovitz, Michelle; Makowsky, Mallory; He, Philip; Kurland, John F; Negro, Alejandra; Sangro, Bruno. (2022). Tremelimumab plus Durvalumab in Unresectable Hepatocellular Carcinoma. NEJM evidence. https://doi.org/10.1056/EVIDoa2100070.
14. Ioannidis, John P. A.. (2005). Why Most Published Research Findings Are False. PLoS Medicine. https://doi.org/10.1371/journal.pmed.0020124.

15. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.

16. Chan, An-Wen; Hróbjartsson, Asbjørn; Haahr, Mette T.; Gøtzsche, Peter C.; Altman, Douglas G.. (2004). Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials. JAMA. https://doi.org/10.1001/jama.291.20.2457.

17. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.

18. 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. https://doi.org/10.1038/sdata.2016.18.

19. Higgins, J. P. T.; Jackson, D; Barrett, J. K.; Lu, G; Ades, A. E.; White, I. R.. (2012). Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. Research Synthesis Methods. https://doi.org/10.1002/jrsm.1044.

20. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.

21. Li, Qinfei; Li, Hong; Ma, Haowei; Chen, Wei. (2026). Efficacy and safety of first-line immunotherapy and targeted therapy in advanced HCC: a network meta-analysis with subgroup analysis based on HBV and HCV infection. Frontiers in Immunology, 17, 1706446. https://doi.org/10.3389/fimmu.2026.1706446.

22. Ren, Zhenggang; Xu, Jianming; Bai, Yuxian; Xu, Aibing; Cang, Shundong; Du, Chengyou; Li, Qiu; Lu, Yinying; Chen, Yajin; Guo, Yabing; Chen, Zhendong; Liu, Baorui; Jia, Weidong; Wu, Jian; Wang, Junye; Shao, Guoliang; Zhang, Bixiang; Shan, Yunfeng; Meng, Zhiqiang; Wu, Jianbing; Gu, Shanzhi; Yang, Wei; Liu, Chao; Shi, Xuetao; Gao, Zhenyuan; Yin, Tao; Cui, Jiuwei; Huang, Ming; Xing, Baocai; Mao, Yilei; Teng, Gaojun; Qin, Yanru; Wang, Jinhai; Xia, Feng; Yin, Guowen; Yang, Yong; Chen, Mingxia; Wang, Yan; Zhou, Hui; Fan, Jia; ORIENT-32 study group, . (2021). Sintilimab plus a bevacizumab biosimilar (IBI305) versus sorafenib in unresectable hepatocellular carcinoma (ORIENT-32): a randomised, open-label, phase 2-3 study. The Lancet. Oncology. https://doi.org/10.1016/S1470-2045(21)00252-7.

23. Robinson, W. S.. (1950). Ecological Correlations and the Behavior of Individuals. American Sociological Review. https://doi.org/10.2307/2087176.

24. Guyatt, Gordon; Oxman, Andrew D.; Akl, Elie A.; Kunz, Regina; Vist, Gunn; Brozek, Jan; Norris, Susan; Falck-Ytter, Yngve; Glasziou, Paul; deBeer, Hans. (2011). GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. Journal of Clinical Epidemiology. https://doi.org/10.1016/j.jclinepi.2010.04.026.

25. Hutton, Brian; Salanti, Georgia; Caldwell, Deborah M.; Chaimani, Anna; Schmid, Christopher H.; Cameron, Chris; Ioannidis, John P.A.; Straus, Sharon; Thorlund, Kristian; Jansen, Jeroen P.; Mulrow, Cynthia; Catalá-López, Ferrán; Gøtzsche, Peter C.; Dickersin, Kay; Boutron, Isabelle; Altman, Douglas G.; Moher, David. (2015). The PRISMA Extension Statement for Reporting of Systematic Reviews Incorporating Network Meta-analyses of Health Care Interventions: Checklist and Explanations. Annals of Internal Medicine. https://doi.org/10.7326/m14-2385.

26. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

27. Zarin, Deborah A.; Tse, Tony; Williams, Rebecca J.; Califf, Robert M.; Ide, Nicholas C.. (2011). The ClinicalTrials.gov Results Database - Update and Key Issues. New England Journal of Medicine. https://doi.org/10.1056/nejmsa1012065.

28. Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

29. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

30. Page, Matthew J, et al. "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews." *BMJ*, 2021, pp. n71. https://doi.org/10.1136/bmj.n71.