Annals of Science and Inquiry
Vol. 2 No. 11 (2026)

How Completed Phase II Cancer Trials Register Endpoints and Their Time Frames

Louis Schulz · Christopher Hughes · Jack R. Watson

Published: November 26, 2026 | Article 1085

Abstract

Registry searches can place trials next to one another even when they describe their primary endpoints and assessment horizons differently. We examined 122 completed phase II registrations with explicit condition labels for biliary tract cancer or pancreatic ductal adenocarcinoma and completion dates between 2010 and 2024. Among 107 biliary tract cancer records, 51 listed a response-related primary endpoint by a broad label rule, compared with 11 of 15 pancreatic ductal adenocarcinoma records. The difference was −25.7 percentage points, with a 95% score interval from −43.9 to 1.3. Every primary-entry time frame was populated in 106 of 107 and 15 of 15 records, respectively. When a numeric duration unit was required in every entry, the counts became 96 and 15. These differences describe documentation and endpoint selection in a public registry cohort. They do not compare clinical outcomes.

How Completed Phase II Cancer Trials Register Endpoints and Their Time Frames

Abstract

Registry searches can place trials next to one another even when they describe their primary endpoints and assessment horizons differently. We examined 122 completed phase II registrations with explicit condition labels for biliary tract cancer or pancreatic ductal adenocarcinoma and completion dates between 2010 and 2024. Among 107 biliary tract cancer records, 51 listed a response-related primary endpoint by a broad label rule, compared with 11 of 15 pancreatic ductal adenocarcinoma records. The difference was −25.7 percentage points, with a 95% score interval from −43.9 to 1.3. Every primary-entry time frame was populated in 106 of 107 and 15 of 15 records, respectively. When a numeric duration unit was required in every entry, the counts became 96 and 15. These differences describe documentation and endpoint selection in a public registry cohort. They do not compare clinical outcomes.

Registry similarity and the primary outcome

The same registry search can identify trials that answer different questions. An entry that gives a primary response endpoint has a different measurement structure from one focused on time to an event. Both may display a filled assessment field. One field can describe when observers return for measurements, another can specify how long each participant contributes to an endpoint, and another can describe the length of follow-up in general terms. A reader narrowing a set of studies to comparable evidence therefore needs more information than a phase label and a check for the presence of a time box in the record. The aim of this study is to separate those two layers for a narrowly defined public trial cohort.

The indexed title of Yi et al. (2024) contains both condition names used to define the two groups in this administrative comparison. One DOI can therefore be relevant to both name searches. It supplies a reproducible bibliographic reason to examine both condition labels and to retain them separately when constructing the registry roster. A publication title fixes the terminology at citation level. The labels and time frames recorded for the individual public registrations answer a different documentation question. The comparison uses the registered text fields and the timing rules described below, with no attempt to recover study-level values from the citation. This distinction also makes the denominator explicit. One observation is a unique registration within the stated group, and multiple entries on that registration can be counted before aggregation.

Choice of an endpoint reflects a study objective and its design. Simon (1989) provides a formulation for planning a two-stage phase II study around a response endpoint. Foster et al. (2020) discuss interpretation of single-arm combination studies and the importance of the outcome definition when a study is compared with historical information. Wittes (2002) explains why design assumptions about an outcome affect sample size and the precision of inference. The statistical distinction between an endpoint and the quantity estimated by comparing outcomes appears in the ICH estimands addendum (International Council for Harmonisation, 2019). These sources make it plausible that similar-looking registry titles can point to different quantitative questions. Our aim is to measure that distinction in the public record as it can actually be read.

Two primary comparisons focus the audit. The first is whether each study lists any response-related primary endpoint. The second is whether every primary endpoint in the study has a time-frame field with a written number followed by a time unit. Both operate at the level of the registration. We also count time-to-event labels, multiple primary entries, and the presence of any nonempty time-frame text. The last

comparison helps identify the gap between recording some description and recording a duration that a simple rule can recognize.

Construction of the registry cohort

We retrieved design records from the public ClinicalTrials.gov API on 23 November 2026 (National Library of Medicine, n.d.-a). Two condition queries targeted biliary tract or gallbladder cancer and cholangiocarcinoma on one side, and pancreatic ductal adenocarcinoma on the other. The request restricted results to completed interventional studies with a completion date between 1 January 2010 and 31 December 2024 and a phase II match. We collected public condition fields, study phase and date, and the measure and time frame of each primary outcome. Only the fields needed for the analysis were retained. Counts and code identify the dated snapshot, with its digest, so the cohort can be reconstructed without relying on a later search ranking or registry update.

The queries returned 169 and 46 result rows. Deduplication by registry identifier produced 210 records. We required that phase II be the sole phase code and that the condition field explicitly name one target indication, using the patterns in the cohort script. Forty-nine records failed the phase-only rule. Thirty-nine lacked a condition-field label in the prespecified definitions. Of the remaining 122 registrations, 107 matched the biliary tract cancer group and 15 the pancreatic ductal adenocarcinoma group. A registration appearing in both condition groups would have been excluded from the two-group comparison. These inclusion rules define a sample of recoverable records. In particular, studies registered under a broad synonym that lacks an explicit qualifying term can fall outside the cohort. Previous treatment, disease stage and response eligibility were not criteria for this design audit. Accordingly, neither group is an estimate of the exact clinical population of the cited trial.

We use the protocol section visible at extraction. The National Library of Medicine (n.d.-b) describes the primary-outcome entry as a measure with associated time information. Zarin et al. (2011) distinguish registration elements from results records. An assessment of endpoint switching would require histories of registry changes and a linkage to the protocol or article, as illustrated by Mathieu et al. (2009). Hartung et al. (2014) compare information released in the registry with matching publications, and Riveros et al. (2013) examine timing and completeness of the two sources. The present cohort instead records the version visible at one extraction date. Its endpoint features describe the public documentation at that date rather than the text originally posted, and only the design section contributes to the comparisons.

Coding endpoint titles separately from assessment text

We used three families of labels in the primary measure title. The time-to-event rule detects terms such as survival or recognized time-to-event abbreviations. The response rule detects terms such as response, disease control or clinical benefit. An entry with a time-to-event match takes precedence over its response match so duration-of-response wording does not by itself count as a response-rate measure. The safety rule detects terms such as safety, toxicity, adverse events and tolerability. Where titles contain more than one family of terms, safety can coexist with an event or response designation. At study level an indicator is positive if any primary entry has the specified label. This design can capture multiple registered objectives without pretending that the categories divide the records into separate populations. Additional strict response coding requires a more specific title phrase or an established abbreviation and is used as a sensitivity analysis.

The title rules adapt the general distinction between response and time-to-event endpoints found in guidance from the U.S. Food and Drug Administration (2018). Response evaluation frameworks in Eisenhauer et al. (2009) and Seymour et al. (2017) illustrate why the presence of a response label tells a

reader little about the full endpoint definition. The rule does not infer a particular set of response criteria, study threshold, statistical estimator or analysis population. Dodd et al. (2018) show how formal outcome domains can aid evidence discovery, and Elghafari et al. (2021) examine assessment of outcome extraction from registry text. Our deliberately small lexical rules are open to inspection in the code. They are designed to count visible wording consistently and have not been validated against expert adjudication. Results for the broad and strict response rules indicate sensitivity to one readily auditable change in coding.

The time-frame rule distinguishes a field with any written content from a field containing an Arabic numeral and hours, days, weeks, months or years. We also flag recurring wording such as “every” or “monthly” without treating that wording as a distinct estimate of duration. A study passes the all-entry screen if every registered primary outcome satisfies it. A study with no primary entry fails the screen by definition and is retained in the main denominator. A complete-case sensitivity analysis excludes that study. The numeric check is a surface feature of the title-independent time field. An entry can pass because it describes an interval between visits rather than a fixed assessment horizon. Conversely, a text description with no number and unit can be meaningful. We therefore report passage of the lexical check and presence of the field separately. Butcher et al. (2022) describe a broader set of outcome information that matters in protocols and published reports, such as the measurement variable and time point used for analysis. Our two text checks cover only specific features that were observable in the selected fields.

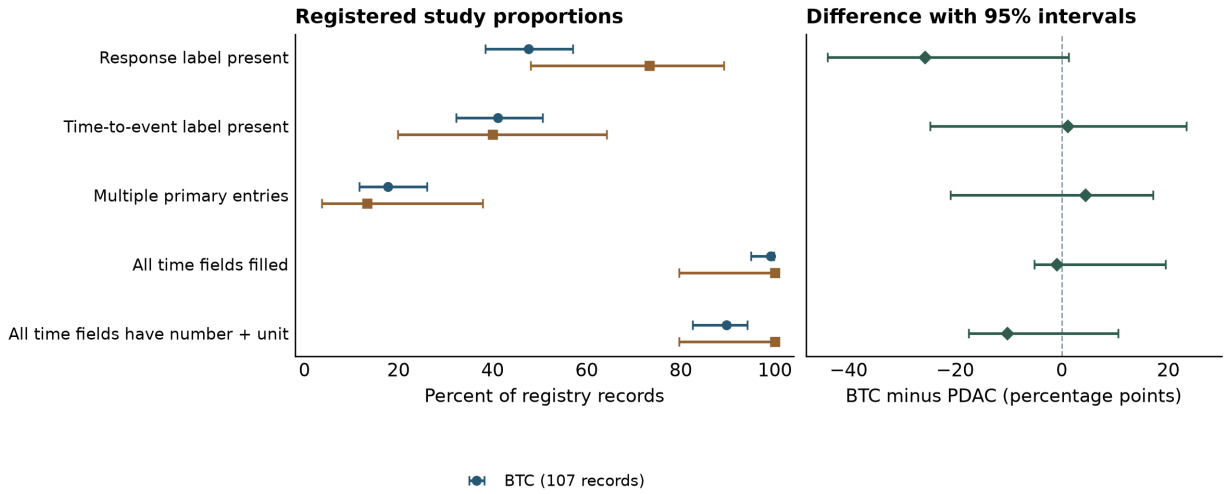

Figure 1. *The left panel presents study-level proportions with Wilson intervals. The right panel presents biliary tract cancer minus pancreatic ductal adenocarcinoma differences with Newcombe score intervals. Entries within a registration contribute to the indicator and do not enter as independent observations.*

Findings by condition group

We identified 132 primary entries in the 107 biliary tract cancer registrations and 18 in the 15 pancreatic ductal adenocarcinoma registrations. One biliary tract cancer record had no primary entry. A response label appeared in 51 of 107 biliary tract cancer registrations and 11 of 15 pancreatic ductal adenocarcinoma registrations. The resulting difference was −25.7 percentage points. The case counts underlying the difference are important because a single change in the smaller group corresponds to about 6.7 percentage points. The strict rule identified 40 of 107 and 10 of 15, respectively. This is a −29.3 point difference under the narrower coding, indicating that the direction is consistent across the two title definitions while their estimates differ.

Table 1. *Study-level endpoint and assessment field indicators.*

Indicator	Biliary tract cancer, n/107	Pancreatic ductal adenocarcinoma, n/15	Difference, percentage points (95% interval)
Multiple primary entries	19	2	4.4 (−20.9 to 17.1)
Any response label, broad rule	51	11	−25.7 (−43.9 to 1.3)
Any response label, strict rule	40	10	−29.3 (−49.4 to −2.6)
Any time-to-event label	44	6	1.1 (−24.7 to 23.4)
Any safety label	11	0	10.3 (−10.6 to 17.5)
All primary entries have time text	106	15	−0.9 (−5.1 to 19.5)
All primary entries have number and time unit	96	15	−10.3 (−17.5 to 10.6)

Table 1 note. *Differences are biliary tract cancer minus pancreatic ductal adenocarcinoma, with 95% Newcombe score intervals. The figures describe registered definitions, and the endpoint categories can overlap across entries within a study.*

The rate of any time-to-event label was similar in the two selected groups: 44 of 107 and six of 15. Safety labels were found in 11 biliary tract cancer records and no pancreatic ductal adenocarcinoma records under the rule. This describes names of primary outcomes, leaving secondary endpoints and results outside the analysis. On the time-field comparison, all entries had text for every study with at least one primary endpoint. The numeric-unit screen nevertheless retained just 96 of those 106 biliary tract cancer registrations and all 15 from the pancreatic ductal adenocarcinoma group. This gives a complete-case difference of −9.4 percentage points compared with −10.3 percentage points in the main denominator. The size of this change shows that the one absent primary record does not account for most of the gap between the two screens.

Wilson (1927) motivates the score interval for an individual proportion. Newcombe (1998) combines score limits for a difference of independent proportions. Brown et al. (2001) and Agresti and Coull (1998) discuss why standard normal intervals for raw proportions can behave poorly for small samples and proportions near zero or one. These concerns apply directly to our smaller group and all-entry indicators. Intervals describe statistical precision under a binomial working model. They do not account for the restrictiveness of the registry search or unknown classification error. No adjustment for multiple comparisons was used. The secondary rows and the strict coding are complementary descriptions rather than independent confirmations of a common claim.

Implications for evidence comparison

The analysis highlights why phase and indication filters should be followed by an endpoint-level check. For example, the difference between the groups in use of response wording cannot establish a difference in response probability. The registry documents what a trial seeks to measure, while clinical response refers to observations in its participants. A similar issue arises when interpreting a short-term endpoint as a surrogate for a later outcome. Buyse et al. (2000) show that validation of a surrogate requires evidence from randomized experiments beyond co-occurrence of names in protocols. This audit's title categories supply a practical description of which definitions a synthesis would have to recover. They cannot establish equivalence of measurement or suitability of a surrogate.

One productive use of the data is to flag records that will need additional retrieval and annotation during a review. For instance, a response endpoint with a recurring assessment schedule could require checking the protocol for the point at which the analysis freezes follow-up. When a record gives multiple primary entries, a reviewer may also need to establish whether those entries represent alternative analyses, separate groups, or a set of jointly assessed outcomes. A trial-level indicator helps prioritize that work. The entry-level file preserves the positions of the original primary outcomes so each annotation can be traced to the appropriate field.

The completeness of public registration records is a separate issue from selective publication. Simes (1986) and De Angelis et al. (2004) describe the value of registration for recognizing trials independently of how their findings are disseminated. Hopewell et al. (2007) examine time to publication, and DeVito et al. (2020) and Goldacre et al. (2018) audit results reporting in two registry systems. The World Wide Web Consortium (2013) describes a provenance model for recording how an observation was derived from a particular source version. Those questions motivate finding and preserving endpoint definitions. Our measured quantities characterize public registrations regardless of how the corresponding trials eventually appeared in journals. The visible snapshot cannot show whether a definition was missing initially, or whether a later revision supplied or changed it. Registry histories would help distinguish those explanations.

This descriptive study uses selected fields from routinely maintained records. Recommendations by von Elm et al. (2007) and Benchimol et al. (2015) support documenting inclusion rules, missing data and the relationship of the analysis to these source records. A follow-up audit could review the time texts manually to distinguish visit schedules, limits and reference events, then link final protocols to dated registry revisions. The release accompanying this study provides the cohort rules, title and time-frame classifiers, trial-level flags, and interval calculations for those checks. For the two indication groups in this sample, counting filled fields gives a more optimistic view of directly recoverable timing than counting numeric durations. Neither number determines whether follow-up was comparable. The information that a synthesis needs remains the measure, the assessment horizon, and the population for which the measure was calculated.

References

Agresti, Alan; Coull, Brent A. (1998). Approximate is Better than “Exact” for Interval Estimation of Binomial Proportions. *The American Statistician*, 52(2), 119–126. DOI.

Benchimol, Eric I.; Smeeth, Liam; Guttmann, Astrid; Harron, Katie; et al. (2015). The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Medicine*, 12(10), e1001885. DOI.

Brown, Lawrence D.; Cai, T. Tony; DasGupta, Anirban (2001). Interval Estimation for a Binomial Proportion. *Statistical Science*, 16(2), 101–133. DOI.

Butcher, Nancy J.; Monsour, Andrea; Mew, Emma J.; Chan, An-Wen; et al. (2022). Guidelines for Reporting Outcomes in Trial Reports. *JAMA*, 328(22), 2252. DOI.

Buyse, M.; Molenberghs, G.; Burzykowski, T.; Renard, D.; Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, 1(1), 49–67. DOI.

De Angelis, Catherine; Drazen, Jeffrey M.; Frizelle, Frank A.; Haug, Charlotte; et al. (2004). Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine*, 351(12), 1250–1251. DOI.

DeVito, Nicholas J; Bacon, Seb; Goldacre, Ben (2020). Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: a cohort study. *The Lancet*, 395(10221), 361–369. DOI.

Dodd, Susanna; Clarke, Mike; Becker, Lorne; Mavergames, Chris; Fish, Rebecca; Williamson, Paula R. (2018). A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery. *Journal of Clinical Epidemiology*, 96, 84–92. DOI.

Eisenhauer, E.A.; Therasse, P.; Bogaerts, J.; Schwartz, L.H.; et al. (2009). New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*, 45(2), 228–247. DOI.

Elghafari, Anas; Finkelstein, Joseph (2021). Automated Identification of Common Disease-Specific Outcomes for Comparative Effectiveness Research Using ClinicalTrials.gov: Algorithm Development and Validation Study. *JMIR Medical Informatics*, 9(2), e18298. DOI.

Foster, Jared C; Freidlin, Boris; Kunos, Charles A; Korn, Edward L (2020). Single-Arm Phase II Trials of Combination Therapies: A Review of the CTEP Experience 2008–2017. *JNCI: Journal of the National Cancer Institute*, 112(2), 128–135. DOI.

Goldacre, Ben; DeVito, Nicholas J; Heneghan, Carl; Irving, Francis; et al. (2018). Compliance with requirement to report results on the EU Clinical Trials Register: cohort study and web resource. *BMJ*, k3218. DOI.

Hartung, Daniel M.; Zarin, Deborah A.; Guise, Jeanne-Marie; McDonagh, Marian; Paynter, Robin; Helfand, Mark (2014). Reporting Discrepancies Between the ClinicalTrials.gov Results Database and Peer-Reviewed Publications. *Annals of Internal Medicine*, 160(7), 477–483. DOI.

Hopewell, Sally; Clarke, Mike J; Stewart, Lesley; Tierney, Jayne (2007). Time to publication for results of clinical trials. *Cochrane Database of Systematic Reviews*, 2010(1). DOI.

International Council for Harmonisation (2019). *E9(R1) addendum on estimands and sensitivity analysis in clinical trials*. Source.

World Wide Web Consortium (2013). *PROV-O: The PROV Ontology*. https://www.w3.org/TR/prov-o/

Yi, Lixia; Pan, Haoqi; Ning, Zhouyu; Xu, Litao; et al. (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. DOI.

Mathieu, Sylvain (2009). Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials. *JAMA*, 302(9), 977. DOI.

National Library of Medicine (n.d.-a). *About the API: ClinicalTrials.gov*. Source.

National Library of Medicine (n.d.-b). *Protocol Registration Data Element Definitions for Interventional and Observational Studies*. Source.

Newcombe, Robert G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, 17(8), 873–890. DOI.

Riveros, Carolina; Dechartres, Agnes; Perrodeau, Elodie; Haneef, Romana; Boutron, Isabelle; Ravaud, Philippe (2013). Timing and Completeness of Trial Results Posted at ClinicalTrials.gov and Published in Journals. *PLoS Medicine*, 10(12), e1001566. DOI.

Seymour, Lesley; Bogaerts, Jan; Perrone, Andrea; Ford, Robert; et al. (2017). iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. *The Lancet Oncology*, 18(3), e143–e152. DOI.

Simes, R J (1986). Publication bias: the case for an international registry of clinical trials. *Journal of Clinical Oncology*, 4(10), 1529–1541. DOI.

Simon, Richard (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, 10(1), 1–10. DOI.

U.S. Food and Drug Administration (2018). *Clinical trial endpoints for the approval of cancer drugs and biologics*. Source.

von Elm, Erik; Altman, Douglas G; Egger, Matthias; Pocock, Stuart J; et al. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *PLoS Medicine*, 4(10), e296. DOI.

Wilson, Edwin B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158), 209–212. DOI.

Wittes, J. (2002). Sample Size Calculations for Randomized Controlled Trials. *Epidemiologic Reviews*, 24(1), 39–53. DOI.

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*, 364(9), 852–860. DOI.