Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Estimating trial-record geography across ascertainment and counting specifications: an empirical methods comparison in metastatic colorectal cancer

Jie Qin
The First Affiliated Hospital of Xi'an Jiaotong University
Tao Liu
The Affiliated Hospital of Yan'an University
Yan Feng
The Affiliated Hospital of Yan'an University
Bei Feng
122407959@qq.com

The Affiliated Hospital of Yan'an University

Research Article

Keywords: clinical trials, colorectal cancer, data provenance, geography, measurement, multinational counting, registries

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10986302/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Geographic analyses of clinical trials often rely on one data source and one rule for allocating multinational records across countries. The extent to which these methodological choices alter geographic estimates is uncertain.

Methods

We conducted a cross-source methodological study by linking metastatic colorectal cancer trial records from ClinicalTrials.gov and a July 2026 TrialTrove/Citeline snapshot to construct their deduplicated union. Geographic estimates used 186 IARC entities; the primary period was 2020–2025. A 3-by-2 design crossed two source-specific ascertainment pipelines and a deduplicated-union specification with fractional counting, which conserved one unit per record, and full counting, which assigned one unit to every documented country. The primary measure was the high-income-country observed-to-expected (O/E) ratio using all-stage colorectal cancer incidence as a burden proxy. We assessed cross-source country-field agreement and used total variation distance (TVD) to compare field, cohort-composition, complete-pipeline, and counting-estimand changes. We calculated percentile intervals from 5,000 paired component bootstrap resamples.

Results

The two source cohorts and the union contained 2,394, 864, and 2,841 eligible records or components across all years; 781, 259, and 891 began in 2020–2025. High-income-country O/E ranged from 1.03 to 1.82 across the six primary specifications. Among 140 matched components with country fields in both sources, 134 (95.7%) had identical normalised country sets and mean Jaccard similarity was 0.984. Changing only the country field produced a fractional TVD of 0.012, compared with composition TVDs of 0.187 and 0.131 and a complete-pipeline TVD of 0.240. Fractional-to-full TVDs ranged from 0.191 to 0.296.

Conclusions

Within the analysed specifications, TVD point estimates for cohort-composition and counting contrasts exceeded those for country-field substitution among matched records. Trial-geography studies should report the complete measurement chain and should not interpret documented countries as activated sites, enrolment, access, or database accuracy.

Background

Clinical-trial registries and commercial intelligence platforms are measurement systems rather than neutral censuses. Their records reflect different search interfaces, ontologies, update cycles, source fields, and curation processes. A geographic analysis can therefore change before statistical modelling begins: the selected source defines the candidate cohort, linkage determines which records represent the same trial, country harmonization determines which labels enter the analysis, and the counting rule determines how multinational records contribute to national estimates.[1–7]

Multinational trials introduce a distinct estimand choice. Under full counting, a trial contributes one unit to every documented country. Under fractional counting, the same trial contributes a total of one unit divided across its documented countries. Full counting describes trial-country presences and gives greater total weight to trials listing many countries; fractional counting describes the distribution of a fixed portfolio of trial records.[8] Both are defensible, but they answer different questions and can alter burden-relative ratios and national concentration.

Previous empirical method-comparison studies have shown that registry fields, publications, administrative sources, and alternative variable definitions may agree for some observations while still producing different analytic cohorts or estimates.[6, 7, 9–12] Geographic clinical-research studies have examined international trial distributions,[13–17] while other work has compared trial activity with disease burden, including in oncology.[18, 19] A country listing does not establish that a site activated, recruited participants, provided access, or received equivalent research investment. Likewise, an observed-to-expected ratio of 1.00 is a proportional reference under a selected burden denominator, not a normative equity threshold.

Metastatic colorectal cancer (mCRC) is a clinically heterogeneous disease.[20, 21] We used it as a case application because the analysed trial portfolio includes disease-specific, advanced-disease, site-specific metastasis, and basket-trial records, creating a demanding setting for source and scope decisions. We aimed to (1) quantify variation in geographic estimates across two source-specific ascertainment pipelines and their deduplicated union under two counting estimands; (2) distinguish country-field variation within matched records from cohort-composition and complete-pipeline variation; and (3) propose an empirically informed reporting chain for reproducible registry-based trial-geography studies. This was a descriptive methodological comparison, not a validation of database accuracy or a study of patient-level access.

Methods

Study design, estimands, and reporting framework

This cross-source methodological study evaluated eight linked stages: data source, source-specific eligibility, cross-source matching, country-field selection, country mapping, counting estimand, geographic estimator, and reporting boundary. The primary 3-by-2 comparison crossed two source-specific ascertainment pipelines (ClinicalTrials.gov and TrialTrove/Citeline) and their deduplicated union

with fractional and full counting. We refer to these as three ascertainment specifications; the union was constructed after matching and was not an independently retrieved third source. Three additional contrasts changed one defined part of the measurement chain while holding specified components fixed: country field, cohort composition, and counting estimand.

The formal methodological objective was specified after exploratory source-assignment work. An August 6, 2026 decision log preceded formal reconstruction of the source-specific cohorts and comparative analyses. Audits corrected union-year precedence and a nonconserving fractional denominator, after which all affected outputs were regenerated. The project was not prospectively registered. Reporting was informed by STROBE and RECORD.[22, 23] The proposed reporting chain organizes the decisions examined here into practical reporting items; it was not developed through consensus or independently validated.

Data sources and setting

The ClinicalTrials.gov pipeline began with 3,463 deduplicated records returned on June 30, 2026 from seven API v2 search terms. The TrialTrove/Citeline pipeline began with 991 nonblank records in a licensed July 2026 snapshot. Each source retained its own membership and fields. Cross-source components were constructed only after matching, and the union preserved source provenance. It was an analytic specification, not a reference standard (Additional file 1: eMethods and eTable 2).

National all-stage colorectal cancer incidence from the International Agency for Research on Cancer (IARC) Global Cancer Observatory, Cancer Today, GLOBOCAN 2024 estimates was the primary burden proxy.[24, 25] The public source was accessed on 13–14 July 2026, and the analysis retained the retrieved country metadata, derivation log, and hashes. The retrieved income-group classification was used as implemented. All-stage incidence is not a direct measure of metastatic prevalence, treatment eligibility, or treatment need.

Record eligibility and source-specific cohorts

Eligible records described interventional anticancer therapy for mCRC, metastatic or unresectable advanced colorectal cancer, site-specific colorectal metastases, or basket trials with explicit colorectal cancer participation. We excluded records centred on nonmetastatic disease, prevention, screening, supportive care without anticancer treatment, noninterventional research, or another cancer without an eligible colorectal cancer cohort.

A strict-core indicator identified records explicitly involving metastatic or unresectable advanced colorectal cancer; the broader scope also allowed metastatic-relevant site-specific and basket records. The primary period included records with documented start dates from January 1, 2020 through December 31, 2025. A separate 1,398-record curated portfolio was not the analytic cohort for this study; where documented, it served only as an upstream field and audit resource.

As presubmission quality control, outcome- and stratum-blinded review covered 2,811 unique ClinicalTrials.gov records in two risk-enriched rounds. After reviewer files were locked, adjudication changed 31 eligibility and 78 scope labels among 78 records; all accepted changes were incorporated before the complete analysis rerun. These nonproportional reviews do not estimate pipeline-wide error rates, and 652 records remained outside triggered review requirements (Additional file 1: eMethods, Presubmission eligibility and scope quality control).

Cross-source matching and union construction

Accepted NCT identifiers were extracted from standalone identifiers in vendor protocol or trial-ID fields and from ClinicalTrials.gov URLs, then normalised, deduplicated, and sorted. Matching used exact accepted identifiers, one documented component-linking harmonization edge, and one human-accepted pair from the retained algorithm-generated nonexact candidate log. Connected components represented matched record groups. Candidate generation did not enumerate every possible cross-source overlap (Additional file 1: eTable 11).

The union retained source membership, original source country fields, and a documented harmonised country-set union. ClinicalTrials.gov start year took precedence when available, with vendor year as fallback. Alternative year-precedence and country-set rules were evaluated as sensitivity analyses. Either-source eligibility and combined source records made the union a constructed specification rather than a gold standard (Additional file 1: eTable 3).

Country fields and mapping

Country labels were deduplicated within records. Agreement analyses compared source-specific normalised country sets and harmonised only semantically equivalent labels. For burden analyses, fields were classified as empty, only non-IARC, mixed mapped and non-IARC, or only IARC-mapped. Non-IARC labels were retained for provenance but received no proxy allocation; mixed records were renormalised across mapped entities.

Exact complete-set agreement required identical normalised country sets. Jaccard similarity quantified overlap among components with country fields in both sources. Country-count differences were calculated as TrialTrove/Citeline minus ClinicalTrials.gov counts for eligible matched components.

Counting estimands and statistical analysis

Under fractional counting, a record with k distinct documented locations mapping to prespecified IARC entities contributed $1/k$ to each mapped entity, conserving one unit per record. Under full counting, it contributed one unit to each mapped entity.

The primary measure was the high-income-country O/E ratio:

O/E = high-income-country share of documented record allocation / high-income-country share of colorectal cancer burden.

Secondary measures were income-group O/E, national Herfindahl-Hirschman Index (HHI), the number of 186 IARC entities with zero documented allocation, and pairwise TVD between national allocation distributions. Higher HHI indicates greater concentration. For national distributions *p* and *q*, TVD was one-half the sum of their absolute country-share differences.

Measurement-chain contrasts

The field contrast held matched components, source representation, and counting rule fixed while changing only the source country field. Composition contrasts compared each complete source cohort with the same dual-field-complete matched component set while holding representation and counting rule fixed; a vendor sensitivity used every matched component with an available vendor field. Counting-estimand contrasts held the ascertainment specification fixed while changing fractional to full counting. A complete-pipeline contrast compared the complete ClinicalTrials.gov and TrialTrove/Citeline distributions. These contrasts were descriptive, were not additive, and were not estimates of database accuracy.

Sensitivity analyses and statistical uncertainty

Sensitivity analyses varied the analysis period, burden denominator, clinical scope, vendor curation, registry status, missing-country assumptions, and union construction. The strict-core clinical-scope analysis was a post hoc sensitivity analysis prioritized after audit. The registry-status sensitivity excluded records documented as not yet recruiting, planned, or withdrawn but did not verify activation or enrolment. Missing-field one-country imputations were stress tests rather than statistical bounds. A World Health Organization International Clinical Trials Registry Platform comparison remained exploratory and was not treated as a reference standard.[26]

The cross-source component was the resampling unit. In 5,000 paired bootstrap replicates using seed 20260810, we retained source membership, eligibility, start years, and both source-specific country fields while recalculating all six primary specifications. Percentile intervals describe record-set resampling within observed components. They do not incorporate systematic undercapture, missed linkage, registry-update error, TrialTrove/Citeline curation error, or uncertainty in the IARC denominator.

Reproducibility and use of AI-assisted tools

The frozen analyses used Python 3.12.13, pandas 2.3.3, NumPy 2.3.5, and Matplotlib 3.10.8. An isolated clean rerun using relative paths reproduced the frozen numerical outputs and figures without changing scientific definitions, cohorts, estimands, or results. Implementation-level checks are reported in Additional file 1: Software and Quality Control.

OpenAI Codex desktop application assisted, under investigator supervision, with language revision, code and data processing, local analysis, nonclinical visualization, interpretation checks, document assembly, consistency checking, and reference assistance. Investigators defined all eligibility criteria, mappings, estimands, and adjudications, reviewed the resulting code and outputs, and retained scientific authority. The authors remain responsible for the analyses, figures, text, and citations. The workflow, human-review controls, and data-protection boundaries are described in Additional file 1: eAppendices 1–2.

Results

Source-specific cohort construction

The ClinicalTrials.gov query yielded 3,463 deduplicated records, of which 2,394 met the common eligibility criteria. Among eligible records, 781 had documented start dates in 2020–2025, 712 had country information, and 707 mapped to at least one IARC entity (Table 1; Fig. 1; Additional file 1: eTable 1). The TrialTrove/Citeline snapshot contained 991 nonblank records, of which 864 were eligible; 259 were in the primary period, 254 had country information, and 253 mapped to at least one IARC entity. The deduplicated eligible union contained 2,841 components across all years and 891 in the primary period; 831 had country information and 825 mapped to at least one IARC entity.

Table 1
Source-specific cohort construction

Specification	Eligible, all years	Strict core	Broader scope, non-core	Started 2020–2025	Country field present	≥ 1 IARC entity	No country field
ClinicalTrials.gov	2394	1870	524	781	712	707	69
TrialTrove/Citeline	864	642	222	259	254	253	5
Deduplicated union	2841	2147	694	891	831	825	60

The union preserves source provenance and is a constructed analytic specification, not a gold standard. The two source cohorts and the union are analytic specifications; the union is not an independent data source. IARC, International Agency for Research on Cancer. Strict-core and broader non-core records are mutually exclusive and together form the eligible broad-scope cohort. Eligible and scope-category counts cover all start years; the remaining columns refer to eligible records starting in 2020–2025.

Among the 891 primary-period union components, 627 were ClinicalTrials.gov-only, 154 were shared, and 110 were TrialTrove/Citeline-only. Strict-core records comprised 71.6%, 84.4%, and 49.1% of these groups, respectively. Country fields were missing in 8.8%, 1.3%, and 2.7%, and records listing more than five mapped entities comprised 8.8%, 3.2%, and 3.6%. These differences established a composition contrast but did not identify why records appeared in one source rather than another (Additional file 1: eTable 5).

Ten eligible matched components had discordant source years. Using ClinicalTrials.gov year precedence rather than the initially executed earliest-nonmissing rule changed primary-period inclusion for two components and minimally changed union O/E estimates.

Cross-source country-field completeness and agreement

Cross-source matching identified 459 components across all eligibility states and 415 eligible in at least one source. In the primary period, 154 matched components were eligible: 140 had country fields in both sources, 12 had only a TrialTrove/Citeline field, and two had neither field (Fig. 2; Additional file 1: eTable 4). Among the 140 dual-field-complete components, 134 (95.7%) had identical normalised country sets and mean Jaccard similarity was 0.984. When all 154 eligible matched components remained in the denominator, exact complete-set agreement was 134 of 154 (87.0%). Across all start years, 349 of 381 (91.6%) eligible dual-field-complete components had exact sets.

Primary 3-by-2 geographic estimates

Across the six broad-scope specifications, high-income-country O/E ranged from 1.03 to 1.82 (Table 2; Fig. 3). Under fractional counting, O/E was 1.44 (95% record-set interval 1.36–1.53) for ClinicalTrials.gov, 1.03 (0.89–1.18) for TrialTrove/Citeline, and 1.35 (1.27–1.44) for the union. Under full counting, the corresponding estimates were 1.82 (1.75–1.88), 1.42 (1.22–1.60), and 1.77 (1.69–1.83). The TrialTrove/Citeline fractional interval included the proportional reference. Fractional allocation totals equalled the numbers of records with at least one IARC-mappable entity: 707, 253, and 825.

Table 2
Primary and strict-core geographic estimates

Specification	Counting	Broad-scope HIC O/E (95% interval)	Strict-core HIC O/E (95% interval)	HHI	Zero-allocation entities
ClinicalTrials.gov	Fractional	1.44 (1.36–1.53)	1.20 (1.10–1.31)	0.228	128
ClinicalTrials.gov	Full	1.82 (1.75–1.88)	1.66 (1.56–1.75)	0.087	128
TrialTrove/Citeline	Fractional	1.03 (0.89–1.18)	0.85 (0.67–1.02)	0.352	149
TrialTrove/Citeline	Full	1.42 (1.22–1.60)	1.11 (0.83–1.37)	0.193	149
Deduplicated union	Fractional	1.35 (1.27–1.44)	1.13 (1.03–1.22)	0.247	127
Deduplicated union	Full	1.77 (1.69–1.83)	1.60 (1.49–1.69)	0.095	127

HIC, high-income country; HHI, Herfindahl-Hirschman Index; O/E, observed-to-expected ratio. Intervals are percentiles from 5000 paired component resamples. O/E of 1.00 is a proportional reference under the selected burden proxy, not a normative equity target. All estimates refer to records starting in 2020–2025 using the incidence burden denominator. HHI and zero-allocation counts refer to the primary broad-scope specification. Strict-core estimates are from the post hoc sensitivity analysis. Zero-allocation counts use 186 IARC entities.

Direct ClinicalTrials.gov-minus-TrialTrove/Citeline O/E differences were 0.41 (0.26–0.56) under fractional counting and 0.39 (0.22–0.59) under full counting. National HHI values under fractional counting were 0.228, 0.352, and 0.247 for ClinicalTrials.gov, TrialTrove/Citeline, and the union; under full counting they were 0.087, 0.193, and 0.095. The respective numbers of IARC entities with zero documented allocation were 128, 149, and 127 (Additional file 1: eTable 6).

Measurement-chain contrasts

The TVD between the complete ClinicalTrials.gov and TrialTrove/Citeline distributions was 0.240 (95% record-set interval 0.198–0.303) under fractional counting and 0.240 (0.204–0.337) under full counting. Among the same 140 dual-field-complete components, changing only the source country field produced TVDs of 0.012 (0.004–0.028) under fractional counting and 0.015 (0.009–0.061) under full counting (Additional file 1: eTable 7).

Comparing each complete source cohort with the dual-field-complete shared set produced fractional composition TVDs of 0.187 (0.143–0.268) for ClinicalTrials.gov and 0.131 (0.090–0.193) for TrialTrove/Citeline. Using all 152 matched components with an available TrialTrove/Citeline field reduced the vendor composition TVD to 0.099 (0.068–0.155), which remained larger than the matched-field contrast.

Within-pipeline changes from fractional to full counting produced TVDs of 0.296 (0.259–0.331), 0.191 (0.125–0.255), and 0.291 (0.254–0.325) for ClinicalTrials.gov, TrialTrove/Citeline, and the union. Corresponding O/E increases were 0.38 (0.32–0.43), 0.39 (0.25–0.51), and 0.41 (0.36–0.47). Records listing more than five mapped countries generated most additional full-counting units despite comprising fewer than 9% of mapped records in each pipeline. Within the analysed specifications, the composition and counting TVD point estimates exceeded the matched-field TVD point estimate (Additional file 1: eTable 7).

Sensitivity and boundary analyses

Including records starting in 2026 changed fractional high-income-country O/E to 1.43, 1.02, and 1.34 for ClinicalTrials.gov, TrialTrove/Citeline, and the union. Replacing incidence with mortality changed these estimates to 1.61, 1.15, and 1.51. In the post hoc strict-core sensitivity analysis, fractional O/E was 1.20 (1.10–1.31), 0.85 (0.67–1.02), and 1.13 (1.03–1.22); the corresponding full-counting estimates were 1.66 (1.56–1.75), 1.11 (0.83–1.37), and 1.60 (1.49–1.69). The strict-core TrialTrove/Citeline point estimate was below 1.00, but its interval included the proportional reference (Additional file 1: eTable 8).

The 69, five, and 60 primary-period records without country fields in ClinicalTrials.gov, TrialTrove/Citeline, and the union, respectively, were examined in documented missing-field stress tests. Only five, one, and six nonempty fields, respectively, contained no IARC-mappable entity. These analyses did not convert missing data into estimates of unobserved activation or enrolment (Additional file 1: eTables 9–10).

Discussion

Principal findings

Geographic estimates for mCRC trial records differed across the analysed measurement-chain specifications. Broad-scope high-income-country O/E ranged from 1.03 to 1.82 across the three ascertainment specifications and two counting estimands. Country-field substitution produced little distributional change within matched dual-field-complete records, whereas cohort composition, the complete ascertainment pipeline, multinational counting, and clinical-scope definition produced larger changes.

These findings do not rank databases or countries. High agreement among linkable records with two available fields indicates that the sources often documented the same normalised country set for those records. It does not establish the completeness of either source, the accuracy of unmatched records, or whether any listed site activated. Complete-pipeline differences may reflect search design, ontology, eligibility, source updating, matching, and cohort composition; the present comparisons did not causally separate these mechanisms.

Relation to previous methodological research

Prior comparisons of trial registries, protocols, and publications have identified both agreement and discrepancies in trial characteristics and reporting.[6, 9, 10] Studies comparing routinely collected data sources similarly show that source-specific availability and coding can alter event capture or precision even when sources overlap.[11] Research on alternative definitions of place has framed the same question more directly: whether a measurement definition changes the estimated association.[12] This study extends that empirical method-comparison approach to trial geography by jointly varying ascertainment pipeline and multinational counting, then separating field substitution from composition-related contrasts.

The small matched-field TVD and larger composition contrasts are therefore compatible rather than contradictory. Two sources can agree closely on country fields for shared, complete records while assembling different complete cohorts. Agreement conditional on successful linkage and field availability should not be generalized to unlinked or incomplete records.

Why the counting estimand matters

Full and fractional counting quantify different constructs. Full counting records every documented trial-country presence and gives more total influence to trials listing many countries. Fractional counting conserves one unit per trial record and estimates how a fixed portfolio is distributed. In the present cohorts, fewer than 9% of mapped records listed more than five countries, but these records generated most additional full-counting units. This weighting difference was consistent with higher high-income-country O/E and lower HHI under full counting in all three ascertainment specifications: large multinational trials contributed additional country presences and distributed weight across more countries.

Neither estimand is universally correct. The choice should follow the research question and be explicit in the title, Methods, Results, and interpretation. Studies of documented country presences may prefer full counting; studies allocating a fixed portfolio of trial records may prefer fractional counting. When conclusions differ across rules, both should be reported rather than selecting the more striking estimate.

Methodological and reporting implications

Table 3 proposes a reporting chain informed by the methodological comparisons in this study. It covers source retrieval, eligibility, matching, country fields, mapping, counting, the burden denominator and uncertainty, and interpretation and reproducibility. The items are intended to make measurement choices explicit; they are not a validated reporting guideline and do not establish transportability to other diseases or data systems. Source-specific and union estimates should remain distinguishable, and a union should not be presented as a gold standard without independent evidence. Research on trial-registry data-sharing statements has also identified ambiguity in stated access conditions, supporting the need for clear descriptions of access arrangements.[27]

Table 3
Proposed reporting chain for registry-based trial-geography studies

Measurement stage	Proposed reporting items	Risk if omitted
Data source	Platform, snapshot or version, retrieval date, full query and update boundary.	Unclear candidate population or temporal comparability.
Eligibility and scope	Source-specific criteria, record flow, review, adjudication and post hoc changes.	Selection or misclassification mistaken for source coverage.
Matching and deduplication	Identifiers, candidate-generation rules, complex components and unresolved overlap.	Duplicate counting or missed links; uncertain comparison set.
Country fields	Original field, normalization, missingness and conditioning on field completeness.	Documentation differences confused with cohort differences.
Geographic mapping	Crosswalk version, unmapped labels and treatment of mixed mapped records.	Hidden exclusions or nonconserving allocations.
Counting estimand	Full or fractional allocation, unit of analysis and definition of k.	Unclear extra weight assigned to multinational records.
Denominator and uncertainty	Burden measure and year, O/E reference, TVD and HHI definitions, resampling unit and paired design.	Ratios mistaken for equity; intervals misread as total uncertainty.
Interpretation and reproducibility	Meaning of country listings, nonadditivity of contrasts, code access and restrictions on reconstruction.	Unsupported claims about activation, enrolment or access; unverifiable reuse.
The proposed items are informed by this empirical case comparison, not by formal consensus or external validation. Full counting measures documented trial-country presence; fractional counting allocates a fixed portfolio of records. Choose the estimand to match the research question and report both when interpretation depends on that choice. Contrasts are not an additive causal decomposition.		

Burden-relative O/E requires equally disciplined language. All-stage colorectal cancer incidence was a transparent proxy in this application, not a direct measure of metastatic treatment need. O/E of 1.00 represented proportional allocation under that proxy, not fairness, adequacy, access, or an ethical target. TVD and paired measurement-chain contrasts provide the more general conclusion: analytic choices can alter an estimated national distribution even within the same clinical domain.

Strengths and limitations

This study used complete source-specific eligible cohorts, preserved source provenance, distinguished original from harmonised country fields, and compared six primary specifications on a common scale. The matched-field, composition, complete-pipeline, and counting contrasts made the changed and fixed

parts of the measurement chain explicit. A component-level paired bootstrap preserved cross-source dependence, and an isolated clean run reproduced all frozen scientific outputs.

Several limitations constrain interpretation. First, the study evaluated two source- and date-specific pipelines and cannot represent every mCRC trial worldwide. Search, ontology, interface, update, screening, and adjudication differences prevented attribution of complete-pipeline contrasts to database coverage alone. Eligibility was partly harmonised through exact-NCT-linked decisions, which may have reduced classification-related differences. TrialTrove/Citeline-field eligibility lacked an accepted independent review whose adjudications were incorporated into the analytic cohort; an exploratory review was not incorporated (Additional file 1: eTable 11).

Second, field-agreement analyses were conditional on linkable records, mostly those sharing NCT identifiers, and on availability of both country fields. Composition contrasts based on dual-field-complete records selected a specific shared subset; the field-available TrialTrove/Citeline sensitivity reduced but did not eliminate the contrast. The bootstrap represented resampling within observed components and did not incorporate missed links, source undercapture, registry-update error, vendor-curation error, or IARC uncertainty. Comparisons of the magnitudes of TVD contrasts are descriptive; paired intervals for the differences between these TVDs were not calculated, and the contrasts use different conditioning sets.

Third, country listings did not measure activated sites, recruitment, enrolment, participant access, research expenditure, or treatment availability. All-stage incidence was an imperfect proxy for mCRC burden, and the retrieved income-group metadata did not support assertion of an unreported reference year. TrialTrove/Citeline licensing prevents public redistribution of row-level records and reconstructive source-difference files. Full restricted-input reconstruction therefore requires authorized access from the original providers. Finally, the objective was refined after exploratory work, and reconstruction corrections and the strict-core sensitivity analysis were post hoc and disclosed.

Conclusions

In this cross-source methodological study, normalised country fields were highly concordant among matched mCRC trial records with fields in both sources, whereas cohort-composition and multinational-counting comparisons had larger TVD point estimates; geographic estimates also varied under the strict-core sensitivity analysis. Registry-based trial-geography studies should preserve source provenance, report the complete measurement chain, and avoid interpreting documented country locations as activated sites, enrolment, access, or database accuracy.

Abbreviations

HHI
Herfindahl-Hirschman Index
IARC

International Agency for Research on Cancer

mCRC

metastatic colorectal cancer

NCT

ClinicalTrials.gov identifier prefix

O/E

observed-to-expected ratio

TVD

total variation distance

Declarations

Ethics approval and consent to participate

The Ethics Committee of Yan'an University Affiliated Hospital determined on July 20, 2026 that this study did not constitute human participants research because it used only trial-registry metadata and aggregate country-level data, involved no interaction or intervention with individuals, and included no individual-level or identifiable private information. Institutional ethics review and informed consent were therefore not required. No reference number was issued. The original institutional record is retained and can be supplied to the journal if requested.

Consent for publication

Not applicable.

Availability of data and materials

Public ClinicalTrials.gov records can be retrieved from ClinicalTrials.gov using the methods described in Additional file 1. The IARC GLOBOCAN 2024 estimates used for the burden denominator were retrieved from the publicly accessible Global Cancer Observatory, Cancer Today, on 13–14 July 2026; investigators may retrieve the public data through IARC's official interface. Analytic code containing no licensed records and nonreconstructive aggregate outputs can be provided by the corresponding author subject to the applicable sharing layer and, where relevant, a data-use agreement. Row-level TrialTrove/Citeline records, account information, complete cross-source linkage materials, and reconstructive source-difference files cannot be redistributed by the authors. The TrialTrove/Citeline snapshot was analysed under institutionally authorized access, and access to the underlying records must occur through Citeline under its applicable terms. The authors will cooperate with editorial or peer-review requests within the licence and applicable legal requirements, but do not assert that Citeline has committed to provide reviewer access. No study-specific public repository or DOI is asserted. Full reconstruction requires authorized access to the restricted TrialTrove/Citeline inputs.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Wu Jieping Medical Foundation (grant 320.6750.2025-06-249). The funder had no role in study design; data acquisition, analysis, or interpretation; manuscript preparation, review, or approval; or the decision to submit the manuscript for publication.

Authors' contributions

JQ: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Validation, Visualization, Writing - Original Draft, Writing - Review & Editing. YF: Conceptualization, Data Curation, Software, Validation, Writing - Original Draft. TL: Conceptualization, Methodology, Supervision, Project Administration, Validation, Writing - Review & Editing. BF: Conceptualization, Investigation, Validation, Writing - Review & Editing. JQ and BF had full access to the data analysed in this study and take responsibility for data integrity and analytic accuracy. JQ performed the statistical analysis.

Acknowledgements

Not applicable.

Clinical trial number

not applicable.

References

1. De Angelis C, Drazen JM, Frizelle FA, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. CMAJ. 2004;171:606–7. 10.1503/cmaj.1041281.
2. Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials.gov results database—update and key issues. N Engl J Med. 2011;364:852–60. 10.1056/NEJMsa1012065.
3. Tse T, Fain KM, Zarin DA. How to avoid common problems when using ClinicalTrials.gov in research: 10 issues to consider. BMJ. 2018;361:k1452. 10.1136/bmj.k1452.
4. Viergever RF, Li K. Trends in global clinical trial registration: an analysis of numbers of registered clinical trials in different parts of the world from 2004 to 2013. BMJ Open. 2015;5:e008932.

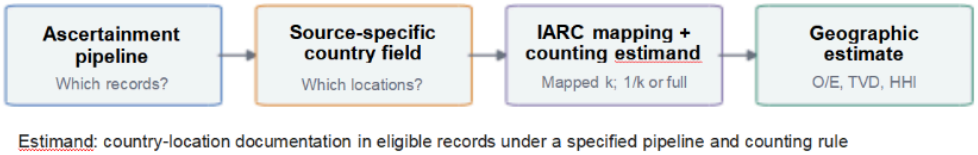
10.1136/bmjopen-2015-008932.

5. World Health Organization. International Clinical Trials Registry Platform. https://www.who.int/tools/clinical-trials-registry-platform. Accessed 21 Aug 2026.
6. Speich B, Gloy VL, Klatte K, et al. Reliability of trial information across registries for trials with multiple registrations: a systematic review. JAMA Netw Open. 2021;4:e2128898. 10.1001/jamanetworkopen.2021.28898.
7. Stergiopoulos S, Getz KA, Blazynski C. Evaluating the completeness of ClinicalTrials.gov. Ther Innov Regul Sci. 2019;53:307–17. 10.1177/2168479018782885.
8. Haeussler C, Rake B. The changing geography of clinical research: a critical analysis of its drivers. Ind Corp Change. 2017;26:285–310. 10.1093/icc/dtx002.
9. Mathieu S, Bouillon-Minois JB, Renard Triché L, et al. Protocol publication rate and comparison between article, registry and protocol in RCTs. BMC Med Res Methodol. 2025;25:31. 10.1186/s12874-025-02471-y.
10. Boyne N, Duke A, Rea J, et al. Discrepancies in safety reporting for chronic back pain clinical trials: an observational study from ClinicalTrials.gov and publications. BMC Med Res Methodol. 2025;25:33. 10.1186/s12874-025-02486-5.
11. Wang A, Seeley AE, Sydes MR, et al. Agreement and utility of coded primary and secondary care data for long-term follow-up of clinical trial outcomes. BMC Med Res Methodol. 2025;25:156. 10.1186/s12874-025-02606-1.
12. Nagy DK, Bresee LC, Eurich DT, Simpson SH. Evaluating methods to define place of residence in Canadian administrative data and the impact on observed associations with all-cause mortality in type 2 diabetes. BMC Med Res Methodol. 2025;25:76. 10.1186/s12874-025-02531-3.
13. Thiers FA, Sinskey AJ, Berndt ER. Trends in the globalization of clinical trials. Nat Rev Drug Discov. 2008;7:13–4. 10.1038/nrd2441.
14. Glickman SW, McHutchison JG, Peterson ED, et al. Ethical and scientific implications of the globalization of clinical research. N Engl J Med. 2009;360:816–23. 10.1056/NEJMsb0803929.
15. Atal I, Trinquart L, Porcher R, Ravaud P. Differential globalization of industry- and non-industry-sponsored clinical trials. PLoS ONE. 2015;10:e0145122. 10.1371/journal.pone.0145122.
16. Lu F, Suda KJ, Marks KE. Geographic location of antineoplastic agent clinical trials conducted in developed and developing countries. Int J Clin Pharm. 2013;35:87–91. 10.1007/s11096-012-9711-6.
17. Payedimarri AB, Mouhssine S, Aljadeeah S, Gaidano G, Ravinetto R. Globalisation of industry-sponsored clinical trials for breast, lung and colon cancer research: trends, threats and opportunities. BMJ Oncol. 2023;2:e000101. 10.1136/bmjonc-2023-000101.
18. Bourgeois FT, Olson KL, Ioannidis JPA, Mandl KD. Association between pediatric clinical trials and global burden of disease. Pediatrics. 2014;133:78–87. 10.1542/peds.2013-2567.
19. Ahuja S, Jainar CJE, Moufarrij S, Dee EC, Jacomina LE. Global disparities in current clinical trials in cervical cancer. JAMA Netw Open. 2026;9:e2626173. 10.1001/jamanetworkopen.2026.26173.

20. Biller LH, Schrag D. Diagnosis and treatment of metastatic colorectal cancer: a review. JAMA. 2021;325:669–85. 10.1001/jama.2021.0106.
21. Cremolini C, Chalabi M, Elez E, et al. Metastatic colorectal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol. 2026;37:759–76. 10.1016/j.annonc.2026.03.005.
22. von Elm E, Altman DG, Egger M, et al. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ. 2007;335:806–8. 10.1136/bmj.39335.541782.AD.
23. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS Med. 2015;12:e1001885. 10.1371/journal.pmed.1001885.
24. International Agency for Research on Cancer. Global Cancer Observatory. Cancer Today. GLOBOCAN 2024 estimates. https://gco.iarc.who.int/today. Accessed 13–14 July 2026.
25. Sung H, Filho AM, Laversanne M, et al. Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide for 34 cancers in 186 countries. CA Cancer J Clin. 2026;76:e70090. 10.3322/caac.70090.
26. World Health Organization. WHO landscape of clinical trials on cancer. https://www.who.int/observatories/global-observatory-on-health-research-and-development/monitoring/clinical-trials-on-cancer. Published Sep 2025. Accessed 21 Aug 2026.
27. Ohmann C, Panagiotopoulou M, Canham S, Felder G, Verde PE. An assessment of the informative value of data sharing statements in clinical trial registries. BMC Med Res Methodol. 2024;24:61. 10.1186/s12874-024-02168-8.

Figures

A



B

	Source/query records	Eligible, all years	Documented start date, 2020-2025	Field present / empty	Mapping all / mixed / only non-IARC
ClinicalTrials.gov	3,463	2,394	781	712 / 69	671 / 36 / 5
TrialTrove/Citeline	991	864	259	254 / 5	244 / 9 / 1
Deduplicated union	Derived	2,841	891	831 / 60	784 / 41 / 6

2020-2025 union components: 627 ClinicalTrials.gov only, 154 in both sources, and 110 TrialTrove/Citeline only. Mapping columns show all mapped / mixed / only non-IARC labels.

Figure 1

Measurement chain and source-specific cohort construction

Panel A shows the measurement chain from source selection through source-specific eligibility, matching, country documentation, IARC mapping, counting estimand, and geographic estimate. Panel B reports query, eligibility, primary-period, country-field, and IARC-mappable counts. The primary-period union comprised 627 ClinicalTrials.gov-only, 154 shared, and 110 TrialTrove/Citeline-only components. The estimand is documented country-location allocation, not site activation, enrolment, or access.

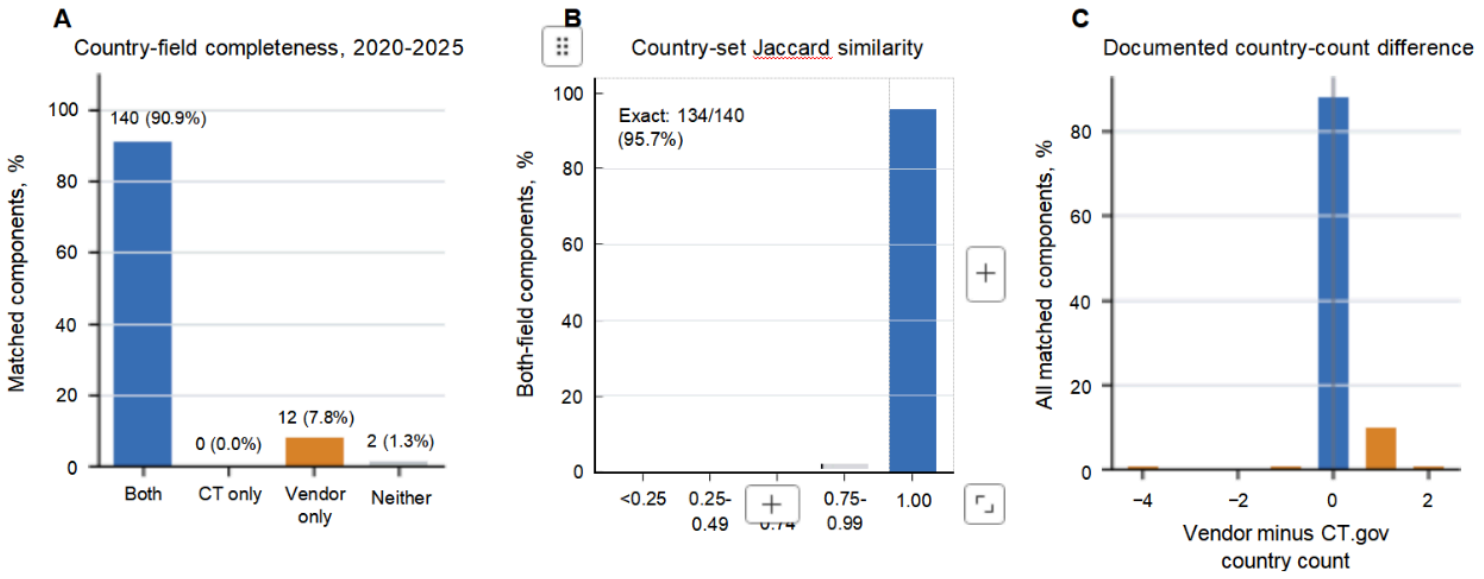


Figure 2

Country-field completeness and cross-source agreement

Among 154 eligible matched components, Panel A shows whether both, one, or neither source supplied a country field. Panel B shows Jaccard similarity among 140 dual-field-complete components; 134 (95.7%) had identical normalised sets and mean Jaccard similarity was 0.984. Panel C shows TrialTrove/Citeline-minus-ClinicalTrials.gov country-count differences among all 154 components. Exact complete-set agreement was 134 of 154 (87.0%) when incomplete fields remained in the denominator.

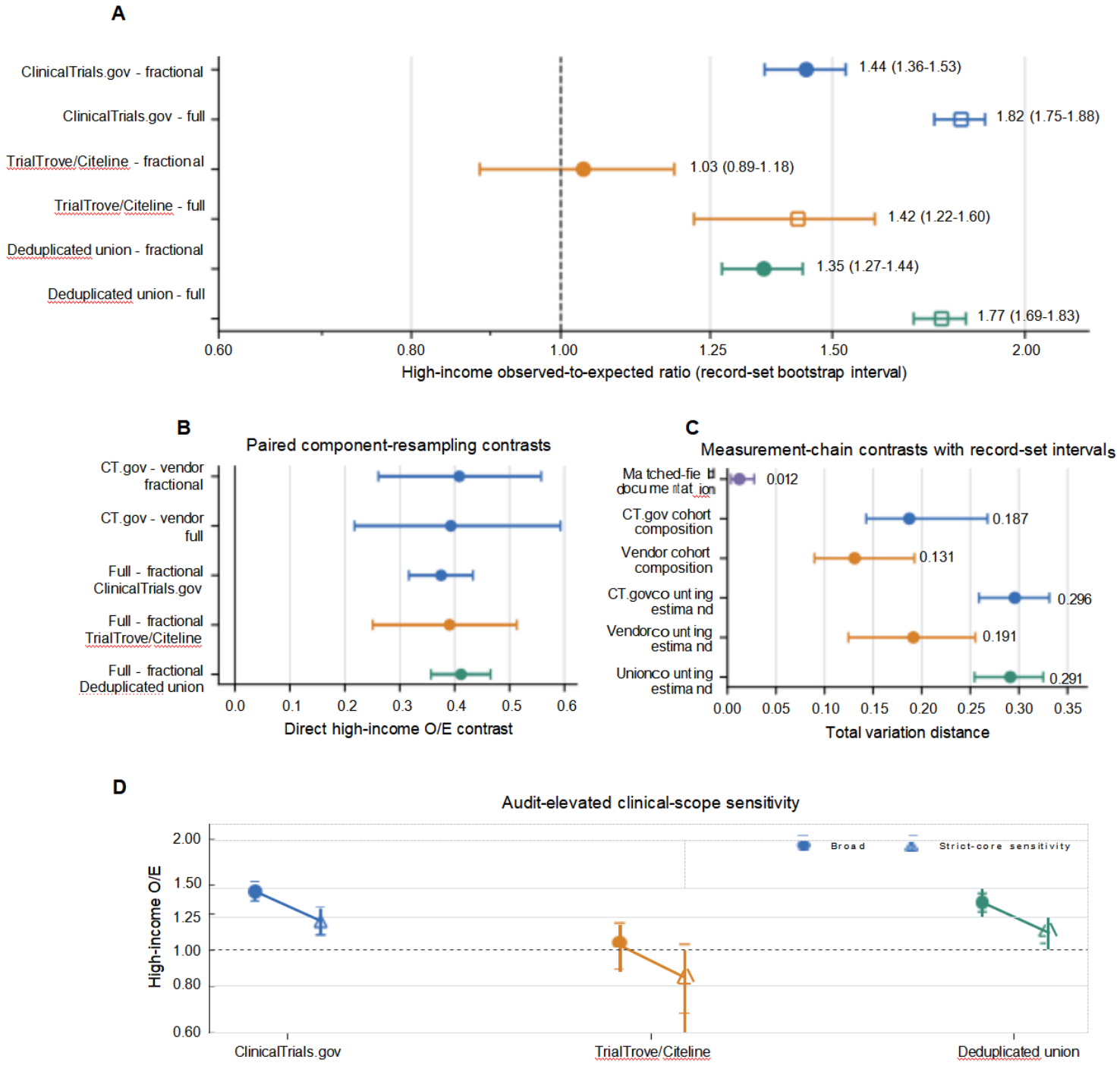


Figure 3

Geographic estimates and measurement-chain contrasts

Panel A presents the primary 3-by-2 high-income-country O/E estimates. Panel B presents direct pipeline and full-minus-fractional O/E contrasts. Panel C presents one matched-field, two cohort-composition, and three counting-estimand TVD contrasts under their specified fixed components and representations. Complete-pipeline TVDs are reported in Results and are not plotted in Panel C. Field and cohort-composition contrasts in Panel C use fractional counting; counting-estimand contrasts compare fractional with full counting within each specification. Panel D contrasts broad scope with the post hoc strict-core sensitivity analysis. Panel C and D labels referring to measurement-chain contrasts and audit-elevated sensitivity denote descriptive, post hoc analyses rather than an additive decomposition. Error bars are the 2.5th and 97.5th percentiles from 5,000 paired component resamples. O/E of 1.00 is a proportional reference, not an equity threshold.