

# A dataset linking 45,988 finished phase 2–3 drug trials on ClinicalTrials.gov (2010–2026) to subsequent FDA approval of the drugs tested

Lars Kullman

[lars@usephaseline.com](mailto:lars@usephaseline.com)

Independent researcher

---

## Data Note

**Keywords:** clinical trials, ClinicalTrials.gov, trial registration, trial termination, protocol amendments, outcome switching, enrollment, landmark analysis, drug development, pharmaceutical industry

**Posted Date:** October 22nd, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11447954/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** The authors declare potential competing interests as follows: The author builds and operates Phaseline ([usephaseline.com](https://usephaseline.com)), a trial-intelligence service that includes a registry-history feature based on the rules described in this manuscript. No funding was received for this work.

---

# Abstract

**Background.** Whether a drug that was tested in a registered trial later reached the market is the outcome that matters for benchmarking trial design and for studies of attrition, yet it is not recorded in any trial registry. Linking registry records to regulatory approvals has so far been done for individual analyses with proprietary pipeline databases.

**Data description.** The dataset contains every interventional phase 2, 2/3 or 3 drug trial on ClinicalTrials.gov that started in 2010 or later and had finished (COMPLETED or TERMINATED) by 18 October 2026: 45,988 trials from every type of sponsor, with the registry's own facts (conditions as MeSH terms, phase, sponsor type, dates, enrollment, randomisation, masking, countries, sites, first primary outcome, posted results, stop reason) and derived labels: drug type, a category for the stop reason, molecular targets, and the FDA outcome of the drugs tested, at two levels. A drug is *novel* for a trial when it had no FDA approval before the trial started; the trial is *approved* when such a drug was later approved for any use (Drugs@FDA), and *approved for the indication* when the trial's condition is named in the approved label. Of 28,083 trials of novel drugs, 23,649 have a decided outcome; 4,865 (20.6 %) led to an approval for any use and 2,601 (11.0 %) to an approval naming the trial's condition. The remaining 4,434 finished too recently to be decided. One CSV file with a data dictionary, the code that builds it, and summary tables are deposited under CC BY 4.0 (data) and MIT (code).

## Background

ClinicalTrials.gov describes trials, not what became of the drugs they tested. Studies of development success rates have therefore relied on commercial pipeline databases [1–3], which are not open and cannot be audited, or on small hand-curated sets. The FDA's own records of approvals (Drugs@FDA) and of current labels are public domain, and the registry's intervention names, once normalised, can be matched to them. This deposit makes that link available for every finished phase 2–3 drug trial since 2010, so that others can use it for benchmarking (how did trials like this one end?), for attrition studies (which design and registry features precede approval?) and for validating prediction methods, without rebuilding the pipeline.

The dataset is the reference set used by Phaseline, a trial-intelligence service built by the author; the matching code is the same as the service runs weekly, and the file is a snapshot of one run. A companion study of registry-history signals and trial termination used the industry subset of the same reference set as its cohort [4].

## Methods

### Trial selection

All studies were read through the ClinicalTrials.gov API (version 2) with the filters interventional, phase 2 or phase 3 (which includes phase 2/3), start date 2010-01-01 or later and overall status COMPLETED or

TERMINATED, on 18 October 2026. A study was kept when at least one intervention was classified as a drug or biological product from its registered type and name, and when its primary completion date (or, failing that, its completion date) was not after the data date. Studies that were withdrawn, suspended, of unknown status or still active are not included. The lead sponsor's registry class gives the sponsor type: INDUSTRY, or everything else (universities, hospitals, NIH and other government, networks and foundations).

## Registry fields

The registry's own content is carried as entered: brief title, lead sponsor, phase, status, the sponsor's reason for stopping, start and primary completion months, enrollment (actual where given, otherwise the last estimate), allocation and masking as binary flags, whether results were posted, the MeSH condition terms and their broader terms that the National Library of Medicine assigns to each record, countries with a site, number of sites, and the first primary outcome measure.

## Derived fields

*Drug type* is assigned by rules from intervention types and names (small molecule; antibody, from the -mab stem and related words; protein or peptide; other, covering cell, gene, nucleic-acid, vaccine and radiopharmaceutical products; mixed when a trial combines types). *Molecular targets* come from the Open Targets drug–mechanism data [5] (CC0) and from WHO INN stems, and are empty when unknown. *Stop category* sorts the sponsor's free-text reason into efficacy, safety, business, enrollment, external, regulatory, other or unstated; rules are applied first, with negations handled, and sentences the rules cannot place were labelled by a large language model reading the sentence (Claude, Anthropic); the free text is kept beside the category.

## FDA outcome

Drug names are taken from each trial's interventions and other names, excluding placebo, vehicle, standard of care and similar entries; salts, hydrates and the four-letter suffixes of biologics are stripped and the names normalised. The openFDA bulk file of Drugs@FDA [6] gives, for every active ingredient and brand name, the date of the first original approval of a new drug application or biologics licence application. A drug is *novel* for a trial when it had no such approval dated before the trial's start; a trial whose drugs were all approved before it started tests new uses of old drugs and is marked accordingly, since its outcome is not in question. A trial with at least one novel drug is *approved for any use* when any of its novel drugs has a first approval on or after the start date; the earliest such approval and the drug's name are recorded. The *indication level* adds a text test: the current label of the approved drug (openFDA drug label, indications and usage section) must contain every distinctive word of at least one of the trial's registered conditions, after removing generic words such as "disease", "chronic" or "advanced" and stemming plurals. A trial of a novel drug with no approval whose primary completion lies less than 36

months before the data date is marked *too recent*, because the interval from primary completion to approval is commonly that long; otherwise it is *not approved*.

## Ethics

Only public registry and regulatory data on trials and organisations were used; no individual-level or investigator data were read, and the registry's contact and investigator names are not part of the dataset. Ethics approval was not required.

## Data description

The file `finished_trials_fda_outcomes.csv` holds one row per trial and 27 columns, documented in the accompanying README; multi-valued fields (MeSH terms, targets, countries) are separated by a vertical bar. Table 1 summarises the trials and Table 2 the outcomes.

Table 1  
The 45,988 trials

| Characteristic                             | Trials | %    |
|--------------------------------------------|--------|------|
| Phase 2                                    | 28,511 | 62.0 |
| Phase 2/3                                  | 2,347  | 5.1  |
| Phase 3                                    | 15,130 | 32.9 |
| Industry lead sponsor                      | 24,290 | 52.8 |
| Academic, government or other lead sponsor | 21,698 | 47.2 |
| Completed                                  | 37,468 | 81.5 |
| Terminated                                 | 8,520  | 18.5 |
| Small molecule                             | 30,679 | 66.7 |
| Antibody                                   | 3,557  | 7.7  |
| Protein or peptide                         | 4,522  | 9.8  |
| Other (cell, gene, nucleic acid, vaccine)  | 1,876  | 4.1  |
| Mixed drug types                           | 5,354  | 11.6 |
| With MeSH condition terms                  | 39,417 | 85.7 |
| With at least one molecular target         | 27,244 | 59.2 |
| With results posted on the registry        | 24,611 | 53.5 |
| Countries represented                      | 167    |      |

Table 2  
FDA outcome of the drugs tested

| Outcome status                              | Trials | Meaning                                                                                    |
|---------------------------------------------|--------|--------------------------------------------------------------------------------------------|
| Approved before start                       | 17,525 | Every drug tested was already FDA-approved when the trial started (new use of an old drug) |
| No drug read                                | 380    | No matchable drug name in the interventions                                                |
| Too recent                                  | 4,434  | Novel drug, no approval, primary completion less than 36 months before the data date       |
| Not approved                                | 18,784 | Novel drug, no approval 36 or more months after primary completion                         |
| Approved, any use                           | 4,865  | A novel drug tested was later approved for some use                                        |
| of which approved for the trial's condition | 2,601  | The condition is named in the approved label's indications                                 |

Among the 23,649 trials of novel drugs with a decided outcome, the share leading to an approval for any use was 17.2% for phase 2 (2,579 of 15,023), 12.0% for phase 2/3 (132 of 1,099) and 28.6% for phase 3 (2,154 of 7,527); at the indication level 7.2%, 6.1% and 19.3%. Industry trials led to an approval in 24.9% of decided cases (3,867 of 15,555) and academic and other trials in 12.3% (998 of 8,094). Among industry phase 3 trials that completed, 36.6% of decided cases (1,862 of 5,088) led to an approval for any use and 26.1% (1,329) to one naming the trial's condition; among terminated industry trials of any phase, 16.2% (442 of 2,724) still led to an approval of the drug, usually in another trial or indication. The interval from primary completion to the first approval of the drug has a median of 10 months (interquartile range – 14 to 33); it is negative for 1,907 of the 4,865 approved trials, where the drug's first approval, often for another indication, preceded the completion of this trial. At the indication level the median is 9 months (– 14 to 28).

Terminated trials give a reason in 7,940 of 8,520 cases. The categories are enrollment 2,829, business or strategic 2,191, efficacy or futility 1,434, unstated 580, other 470, external 444, safety 439 and regulatory 133.

## Limitations

The outcome is a drug's approval, not this trial's contribution to it: a drug approved on the strength of another trial, or for another indication, counts as approved for every trial of a novel drug started before that approval, and the indication-level test narrows but does not remove this. Products regulated by the FDA's Center for Biologics Evaluation and Research (vaccines, blood products, cell and gene therapies) are not in Drugs@FDA and mostly fall under "no drug read" or, when combined with a small molecule, are matched on the comparator; approvals by other regulators are not captured, so a drug approved only in Europe or Japan is "not approved" here. Name matching is by normalised string; code names that never became a non-proprietary or brand name are treated as not approved, and a drug approved under a different salt or formulation may be missed or, rarely, over-matched. The indication test uses the current

label rather than the label at approval. The registry fields are as sponsors entered them, with the usual placeholders and errors; and because trials that stop finish early, the most recent start years are over-represented among terminated trials. The stop categories describe what was written, not what happened. Finally, the file is a snapshot: both the registry and Drugs@FDA change weekly, and a trial marked too recent will be decided in a later release.

## Reuse

The data support, among other uses, benchmarking a planned trial against finished trials with the same conditions, phase and sponsor type; estimating phase-transition rates from an open source rather than a proprietary database, with the caveats above; studying which registry-documented features (size, randomisation, masking, sponsor type, drug type, target) are associated with later approval; and validating prediction methods on a defined cohort with a defined outcome. The code reproduces the file from the registry and openFDA, and future releases will carry the same columns.

## Declarations

## Competing interests

The author builds and operates Phaseline (usephaseline.com), a trial-intelligence service that uses this reference set in its design-check feature and in the validation of its approval-likelihood estimate. No funding was received for this work.

## Data availability

The CSV file (also gzipped), the README with the data dictionary and matching rules, the summary tables and the code (history.py, fda\_outcome.py, dataset\_export.py) are deposited on Zenodo under CC BY 4.0 for the compilation and derived columns and MIT for the code; the registry and FDA content is public domain: doi:10.5281/zenodo.23514302.

## Use of a large language model

The code and the first draft of this paper were produced with the assistance of a large language model (Claude, Anthropic), working under the author's direction; the same model labelled the stop-reason sentences the rules could not place, as described in Methods. The author specified the design, reviewed and tested the code against registry and FDA records, verified the figures in the tables against the deposited file, and edited the text. The model is not an author.

## References

1. Wong CH, Siah KW, Lo AW (2019) Estimation of clinical trial success rates and related parameters. *Biostatistics* 20(2):273–286
2. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J (2014) Clinical development success rates for investigational drugs. *Nat Biotechnol* 32(1):40–51
3. Zhou Y, Zhang Y, Xu H et al (2025) Dynamic clinical trial success rates for drugs in the 21st century. *Nat Commun* 16:9537
4. Kullman L (2026) Enrollment cuts and primary-outcome changes, not completion-date delays, predict the termination of industry phase 2–3 drug trials: a landmark analysis of ClinicalTrials.gov version histories. Preprint, Data and code: 10.5281/zenodo.23457680
5. Ochoa D et al (2023) The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Res* 51(D1):D1353–D1359
6. U.S. Food and Drug Administration. Drugs@FDA and drug label datasets, openFDA. [open.fda.gov](https://open.fda.gov)