Article

Evaluating eligibility criteria of oncology trials using real-world data and AI

https://doi.org/10.1038/s41586-021-03430-5

Received: 24 August 2020

Accepted: 8 March 2021

Published online: 7 April 2021

Check for updates

Ruishan Liu[1], Shemra Rizzo[2], Samuel Whipple[2], Navdeep Pal[2], Arturo Lopez Pineda[2], Michael Lu[2], Brandon Arnieri[2], Ying Lu[3], William Capra[2], Ryan Copping[2]✉ & James Zou[1,3,4,5]✉

There is a growing focus on making clinical trials more inclusive but the design of trial eligibility criteria remains challenging[1–3]. Here we systematically evaluate the effect of different eligibility criteria on cancer trial populations and outcomes with real-world data using the computational framework of Trial Pathfinder. We apply Trial Pathfinder to emulate completed trials of advanced non-small-cell lung cancer using data from a nationwide database of electronic health records comprising 61,094 patients with advanced non-small-cell lung cancer. Our analyses reveal that many common criteria, including exclusions based on several laboratory values, had a minimal effect on the trial hazard ratios. When we used a data-driven approach to broaden restrictive criteria, the pool of eligible patients more than doubled on average and the hazard ratio of the overall survival decreased by an average of 0.05. This suggests that many patients who were not eligible under the original trial criteria could potentially benefit from the treatments. We further support our findings through analyses of other types of cancer and patient-safety data from diverse clinical trials. Our data-driven methodology for evaluating eligibility criteria can facilitate the design of more-inclusive trials while maintaining safeguards for patient safety.

Overly restrictive, and sometimes poorly justified[1], eligibility criteria are a key barrier that leads to low enrolment in clinical trials[2]. For example, around 80% of patients with advanced non-small-cell lung cancer (aNSCLC) did not meet the criteria of the analysed trials[3]. As a result, 86% of clinical trials failed to complete their recruitment within the targeted time[4]. The US National Cancer Institute concluded that eligibility criteria arbitrarily eliminate patients and should be simplified and broadened[5,6]. The US Food and Drug Administration has also emphasized that certain populations are usually excluded from clinical trials without solid clinical justification. Restrictive trials do not fully capture the efficacy and safety of the drug in the populations that will use the drug after approval[1].

There is therefore a great need to have faster trial accrual and better generalizability, with data-driven eligibility criteria[7–10]. However, how to broaden eligibility remains a major challenge. Even trials with similar mechanisms that target the same disease often use different eligibility criteria, possibly owing to legacy protocols. Some eligibility criteria are included to reduce the risks of severe toxicity adverse events, which is a critical consideration[10]. In an evaluation by the American Society of Clinical Oncology, 56% of surveyed clinicians agreed that some criteria are too stringent and harm the trial, but no agreement could be reached on the removal of specific criteria, given the available data[9].

Data-driven algorithms combined with real-world data can potentially improve several aspects of clinical trials[11–13]. Artificial intelligence can screen patients that meet eligibility[14–16], predict which patients are more likely to enrol in trials[17,18] and extract features from electronic health records (EHRs)[19–21]. Several studies have introduced approaches to quantify the difference between the study samples of a clinical trial and the target population that can use the treatment[22,23]. Recent research also used EHR data to evaluate how different eligibility criteria can affect the number of adverse events associated with COVID-19 that are observed in the selected cohort[24]. Our study differs from these studies in that we focus on evaluating the effect of relaxing specific eligibility criteria on treatment efficacy and cohort size in a real-world population. The Flatiron Health database that we use has effectively been used to analyse outcomes of patients with lung cancer after immunotherapies[25,26].

Overview of Trial Pathfinder

We developed Trial Pathfinder as a framework that integrates real-world data and systematically analyses the hazard ratio of the overall survival for cohorts that are defined by different eligibility criteria (Fig. 1). In the first step—trial emulation—we selected individuals in the real-world dataset who met the available eligibility criteria as originally published in the clinical trial protocol. The eligibility criteria were extracted from free text and encoded into programmatic logic statements (Methods). We assigned the selected patients to the treatment groups that were consistent with their treatment records in the Flatiron database. We used the inverse probability of treatment weighting to adjust for baseline confounding factors and to emulate randomization. We then performed survival analysis for the emulated trials using the hazard

[1]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [2]Genentech, South San Francisco, CA, USA. [3]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [4]Department of Computer Science, Stanford University, Stanford, CA, USA. [5]Chan Zuckerberg Biohub, San Francisco, CA, USA. ✉e-mail: copping.ryan@gene.com; jamesz@stanford.edu

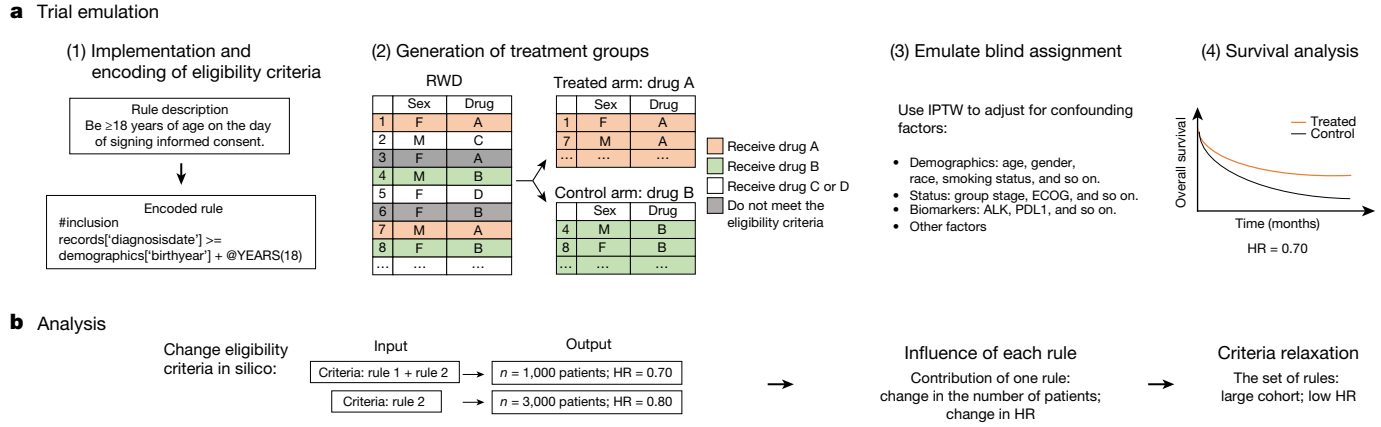


Fig. 1 | Trial Pathfinder workflow and applications. a, Trial Pathfinder takes as input the real-world dataset and the target trial protocol (treatments and eligibility criteria). It programmatically encodes the eligibility criteria and performs trial emulation using propensity score weighting. It then performs a survival analysis on the emulated treatment groups, and reports both the number of eligible patients and the resulting hazard ratio. **b**, Combining an importance analysis of the automated criteria with the Shapley value, Trial Pathfinder evaluates individual criteria and derives a data-driven set of criteria that expands the pool of eligible patients without reducing the effect size. This can guide the design of trials. ALK, anaplastic lymphoma kinase; ECOG, Eastern Cooperative Oncology Group; HR, hazard ratio; IPTW, inverse probability of treatment weighting; PDL1, programmed death ligand 1; RWD, real-world data.

ratio of overall survival as the outcome. The Trial Pathfinder emulation framework makes it possible to systematically vary the eligibility criteria in silico and quantify how the hazard ratio of overall survival changes with different combinations of criteria.

Real-world data and trial emulation

This retrospective study used the Flatiron Health EHR-derived database (https://flatiron.com/real-world-evidence), which includes de-identified data from approximately 280 cancer clinics in the USA[27]. Longitudinal de-identified patient-level data included structured and unstructured data curated from the EHRs. We focused on analysing aNSCLC trials because they have the largest number of patients in the Flatiron Health database, comprising 61,094 patients with aNSCLC. Starting from all of the phase-III aNSCLC trials on ClinicalTrials.gov (queried on 8 November 2019), we filtered for trials that had available trial protocols and had at least 250 patients in each arm in the Flatiron Health dataset who matched the description of the patients in the trials. This resulted in 10 completed aNSCLC trials sponsored by diverse companies that we analysed using Trial Pathfinder (Extended Data Fig. 1 and Extended Data Table 1). Four trials are for first-line treatment and six are for second-line treatment.

Using the Flatiron Health data, we encoded commonly used eligibility criteria based on patient characteristics, diagnoses, laboratory values, biomarkers and previous treatments (Supplementary Table 1). There is substantial heterogeneity in which eligibility criteria were used in each aNSCLC trial, even though they all have the same mechanism of action as checkpoint inhibitors (Extended Data Fig. 2). For example, one trial excluded patients on the basis of albumin and lymphocyte levels, whereas the other nine trials did not. This motivated us to investigate the influence that each inclusion or exclusion criterion had on the real-world population.

Effects of the eligibility criteria

For each aNSCLC trial, we first selected all of the patients with aNSCLC in the Flatiron Health database who have taken the treatment or control drugs in the corresponding line of therapy. On average, 5,167 patients were identified for each trial (Table 1). The hazard ratio of the overall survival was estimated with propensity scores to control for differences between groups (Extended Data Fig. 3). This analysis corresponds to the hypothetical setting in which we fully relax the eligibility criteria.

We next emulated each aNSCLC trial using all of the original protocol criteria that can be encoded in the Flatiron database. The number of patients in the Flatiron database who met all of the eligibility criteria of the trial, along with their emulation hazard ratio of the overall survival, is shown in Table 1. The emulation results are broadly consistent with those of the original randomized trials. On average, only 30% of the patients in the Flatiron database who have taken the drugs tested in the trial actually met the trial eligibility criteria. Moreover, across the trials, the hazard ratio of the full patient population is comparable to, and sometimes smaller than, the hazard ratio of the subset of the patients who met the eligibility criteria (Supplementary Table 2). This suggests that many patients who were excluded by the restrictive eligibility criteria can also potentially benefit from the treatment in the trial.

The above findings motivated us to quantify how each inclusion/exclusion criterion affects the number of eligible patients and the trial outcome. The latter is particularly challenging because the effect of each inclusion/exclusion rule on the hazard ratio depends on which other inclusion/exclusion rules are used to select patients. To estimate this effect systematically, for each aNSCLC trial, we simulated thousands of synthetic cohorts using the Flatiron database under different random combinations of inclusion/exclusion criteria and estimated the hazard ratio of the overall survival for each cohort. We then used the Shapley value[28], an attribution method used in artificial intelligence, to summarize the influence of each criterion. The Shapley value is a weighted average of the effect on the hazard ratio of adding each criterion to different sets of inclusion/exclusion rules (see Methods for details). In our setting, a Shapley value smaller than zero suggests that including the criterion improves the efficacy of the trial and decreases the hazard ratio.

Figure 2 shows the Shapley values for each eligibility criterion estimated with an efficient Monte Carlo algorithm (Methods and Extended Data Fig. 4). Shapley values close to zero (shown in white) correspond to eligibility criteria that had no effect on the hazard ratio of the overall survival. Criteria with beneficial effects (that is, including the criterion would decrease the hazard ratio of the overall survival on average) are shown in blue and detrimental effects (that is, including the criterion would increase the hazard ratio of the overall survival on average) are shown in red. Figure 2 also shows the decrease in the number of eligible patients when each criterion was applied (see Supplementary Tables 3, 4 for the exact numbers).

Our analysis suggests that several commonly used inclusion/exclusion criteria do not substantially affect the hazard ratio of the overall

Table 1 | Comparisons of eligibility criteria

Trial name	Original trial criteria			Fully relaxed criteria		Data-driven criteria		
	No. of criteria	No. of patients	HR	No. of patients	HR	No. of criteria	No. of patients	HR
FLAURA	10	2,277	0.81	3,819	0.82	4	2,546	0.75
LUX8	11	129	0.65	1,350	0.81	5	141	0.58
Checkmate017	17	523	0.67	4,900	0.71	7	4,085	0.71
Checkmate057	19	792	0.75	4,900	0.71	9	2,594	0.66
Checkmate078	18	1,509	0.74	4,900	0.71	9	3,348	0.68
Keynote010	13	806	0.56	1,950	0.51	1	1,948	0.51
Keynote189	15	4,066	0.88	8,818	0.94	7	4,595	0.85
Keynote407	13	2,031	1.13	10,437	1.07	4	9,173	1.04
BEYOND	12	2,902	1.09	9,310	1.14	4	3,043	1.08
OAK	19	493	0.88	1,288	0.87	6	620	0.80
Average	15	1,553	0.82	5,167	0.83	6	3,209	0.77

The number of inclusion/exclusion criteria, the number of eligible patients and the hazard ratio of the overall survival of emulated aNSCLC trials with eligibility criteria under three scenarios: the original criteria used in the trial, fully relaxed criteria and data-driven criteria. The fully relaxed criteria correspond to evaluating the hazard ratio of the overall survival of all of the patients in the Flatiron database who took the treatments in the relevant line of therapy. The data-driven criteria were selected by Shapley values. HR, hazard ratio.

survival of a trial or potentially reduce the efficacy of the trial. These criteria include conditions analysed by laboratory tests (blood pressure, albumin levels, counts of lymphocytes or neutrophils, or alanine aminotransferase (ALT), alkaline phosphatase (ALP) and aspartate aminotransferase (AST) levels) and previous treatments (ALK, PDL1, EGFR and CYP34A therapies, systemic or antineoplastic therapies). These inclusion/exclusion criteria can be restrictive; for example, requiring the lymphocyte count to be greater than 500 per μl excludes 6.3% of the patients on average. Moreover, patients excluded by these criteria benefit from the treatments of the trial to a extent similar to that of patients who met the criteria, as reflected in a Shapley value close to zero (Fig. 2).

Relaxing trial eligibility criteria

The results above show that it is promising to explore the benefits and trade-offs of relaxing standard eligibility criteria. We investigate this by keeping for each trial only the subset of the criteria that Trial Pathfinder identified to decrease the hazard ratio of the trial (that is, with a Shapley value less than zero) and relax the remaining restrictions. We denote this subset the 'data-driven criteria' (Supplementary Table 5). The set of data-driven criteria removes nine inclusion/exclusion rules on average. The hazard ratio of the overall survival had an average reduction of 0.05 compared with using the full eligibility criteria, and the number of eligible patients increased from 1,553 to 3,209 on average, an 107% increase (Table 1 and Extended Data Fig. 5).

Relaxing restrictive eligibility criteria has the important benefit of making clinical trials more inclusive for diverse populations (Supplementary Tables 6–8). The patients who would be excluded by the original trial criteria but would be eligible in the relaxed rules tend to include more women and more patients older than 75 years. Detailed comparisons of patient characteristics between the original trial cohort and our emulations are shown in Supplementary Tables 9–18.

Additional validations

We performed several analyses to support the robustness of our results. In addition to using overall survival as the end point, we repeated all of the analyses for each trial using progression-free survival (Extended Data Table 2). To assess the robustness of our findings in light of the recent shift towards immunotherapies, we ran an analysis in which the data-driven criteria were applied to patients who received treatment between February 2017 and February 2020 (Supplementary Table 19). To assess the representativeness of our findings, we stratified our patient populations on the basis of geographical regions in the USA and the types of insurance plan (Supplementary Tables 20–28). We also applied Trial Pathfinder to 9,439 patients with aNSCLC who received Foundation Medicine genomic tests (Supplementary Tables 29–31). The results of all of these analyses are consistent with our primary findings.

Our primary analyses focused on aNSCLC trials because this cancer type had the most patients in the Flatiron Health database. To investigate the generalizability of Trial Pathfinder to other types of cancer, we identified three additional trials in colorectal cancer, advanced melanoma and metastatic breast cancer with available trial protocols that can be encoded in the Flatiron database (Supplementary Table 32). In all three types of cancer, we found that the original trial criteria were overly restrictive. The data-driven criteria selected by Trial Pathfinder substantially increased the patient population (53% increase on average) while achieving a lower hazard ratio of the overall survival than the original trial criteria (a decrease of 0.13 in the hazard ratio of the overall survival on average) (Extended Data Table 3 and Supplementary Table 33).

Broadening the thresholds of laboratory tests

To more directly assess the effects on safety when broadening eligibility criteria, we analysed the follow-up and evaluation of toxicity for 22 completed Roche oncology trials, which combined comprised 11,602 patients. We found substantial heterogeneity in the eligibility criteria across these trials (Supplementary Table 34). Even trials that targeted the same cancer, in the same phase, and that involved treatments of similar mechanisms used a number of different thresholds of laboratory values to exclude patients. Across aNSCLC, advanced melanoma, metastatic breast cancer and follicular lymphoma, trials with more relaxed thresholds of laboratory values for eligibility did not have more treatment withdrawals due to adverse events than trials with more stringent eligibility thresholds (Supplementary Table 35). This supports our finding that we can potentially broaden several common laboratory-based eligibilities—levels of bilirubin, platelets, haemoglobin and ALP—to align with successful trials that already use these relaxed thresholds without increasing the toxicity risks for the patients.

We further support our findings by analysing abstracted toxicity data in a cohort of 1,000 patients with aNSCLC from the Flatiron database. No significant difference in their baseline laboratory values at the start of treatment were found when comparing patients who reported toxicity-related adverse events during the course of treatment with

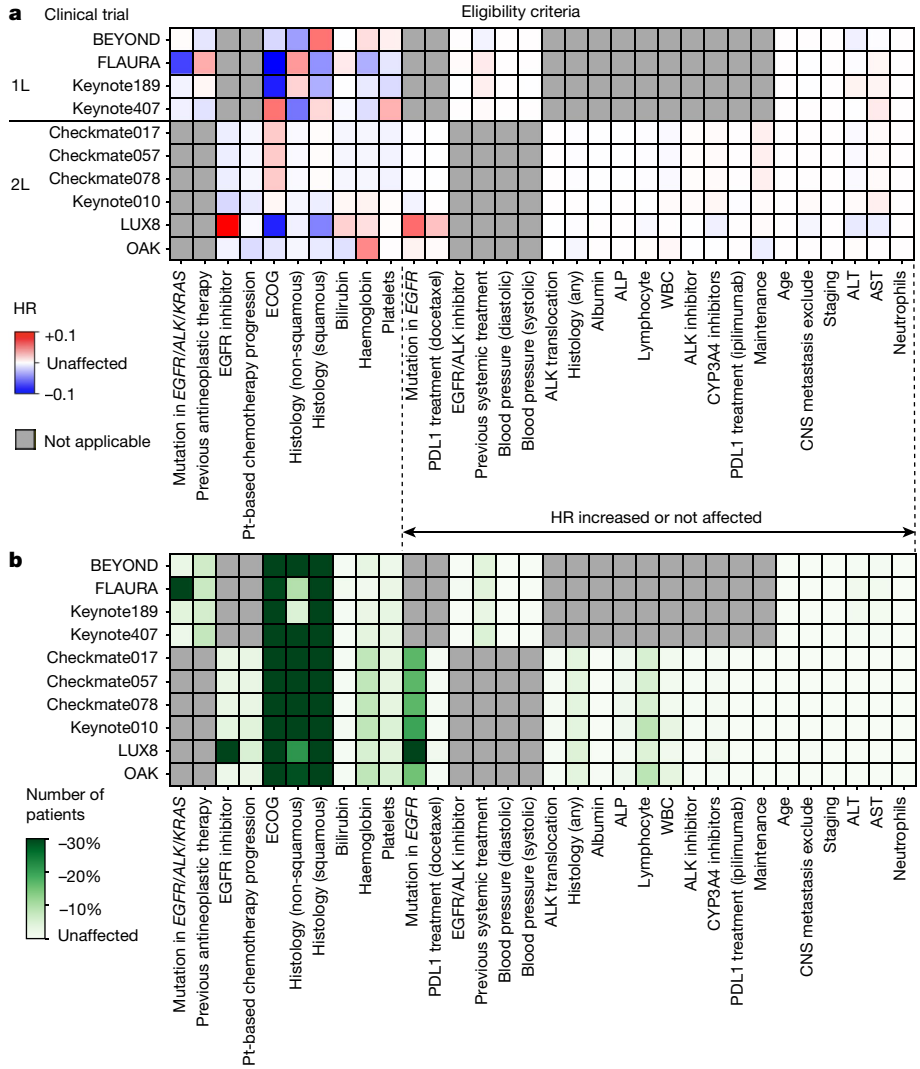


Fig. 2 | Influences of individual eligibility rules. a, **b**, Shapley values of the hazard ratio of overall survival (**a**) and changes in the number of eligible patients (**b**) are shown across different aNSCLC trials and eligibility criteria. **a**, Red, inclusion of the criterion increases the hazard ratio; blue, the criterion decreases the hazard ratio when included, on average. **b**, The fraction of patients who would be excluded by each criterion in every trial is shown. 1L, first line of therapy; 2L, second line of therapy; CNS, central nervous system; Pt, platinum; WBC, white blood cell count. The 'CNS metastasis exclude' criterion means that patients with CNS metastases are excluded.

patients who did not (Extended Data Fig. 6). This reinforces our finding that the broader eligibility thresholds for laboratory tests are feasible from a safety perspective. Furthermore, Extended Data Fig. 7 shows that relaxing the cut-off threshold for the levels of bilirubin, haemoglobin, platelets and ALP within the range of thresholds used in trials (Supplementary Table 35) does not significantly increase the hazard ratio of the overall survival in the Flatiron database and can make trials more inclusive (Supplementary Tables 36, 37).

Discussion

Overly restrictive eligibility criteria limit the access of patients to potentially beneficial treatments. Our findings suggest that it is particularly promising to standardize and potentially broaden several eligibility criteria based on cut-offs for bilirubin, platelets, haemoglobin and ALP values. Recent oncology trials often used different cut-off thresholds for these laboratory tests to exclude patients. We found that across different types of cancer, trials with more relaxed thresholds of laboratory values for eligibility did not have more treatment withdrawals due to adverse events compared with trials with more stringent eligibility thresholds. Together with our findings on the Flatiron data, this suggests that standardizing the eligibility criteria to align with successful trials within the same therapy class that used more relaxed laboratory thresholds could be a good approach to enhance inclusivity.

Because the real-world population can differ from the clinical trial samples, our study aim was not to replicate the original trial results using the Flatiron database. Instead, we investigated how varying the eligibility rules affects the proportion of the real-world population that would be eligible for the trial. Our data-driven evaluation of eligibility criteria should be interpreted as one factor among several that can assist clinical trial specialists in their designs. In each trial, there could be drug-specific nuances, and our hope is that by combining our recommendations with their expertise, trial designers can arrive at more-informed criteria. Currently, longitudinal real-world data with robust outcomes are more limited for diseases other than cancer, which can have more complex end points. There will be opportunities to extend this work outside of oncology as additional high-quality data become available.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information,

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03430-5.

1. Food and Drug Administration. *Enhancing the Diversity of Clinical Trial Populations — Eligibility Criteria, Enrollment Practices, and Trial Designs Guidance for Industry*. https://www.fda.gov/regulatory-information/ search-fda-guidance-documents/enhancing-diversity-clinical-trial-populations- eligibility-criteria-enrollment-practices-and-trial (2020).
2. Van Spall, H. G., Toren, A., Kiss, A. & Fowler, R. A. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *J. Am. Med. Assoc.* **297**, 1233–1240 (2007).
3. Fehrenbacher, L., Ackerson, L. & Somkin, C. Randomized clinical trial eligibility rates for chemotherapy (CT) and antiangiogenic therapy (AAT) in a population-based cohort of newly diagnosed non-small cell lung cancer (NSCLC) patients. *J. Clin. Oncol.* **27**, 6538 (2009).
4. Huang, G. D. et al. Clinical trials recruitment planning: a proposed framework from the Clinical Trials Transformation Initiative. *Contemp. Clin. Trials* **66**, 74–79 (2018).
5. National Cancer Institute. *Report of the National Cancer Institute Clinical Trials Program Review Group*. http://deainfo.nci.nih.gov/advisory/bsa/bsa_program/bsactprgmin.pdf (2017).
6. Mendelsohn, J. et al. *A National Cancer Clinical Trials System for the 21st Century: reinvigorating the NCI Cooperative Group Program* (National Academies Press, 2010).
7. George, S. L. Reducing patient eligibility criteria in cancer clinical trials. *J. Clin. Oncol.* **14**, 1364–1370 (1996).
8. Fuks, A. et al. A study in contrasts: eligibility criteria in a twenty-year sample of NSABP and POG clinical trials. *J. Clin. Epidemiol.* **51**, 69–79 (1998).
9. Kim, E. S. et al. Modernizing eligibility criteria for molecularly driven trials. *J. Clin. Oncol.* **33**, 2815–2820 (2015).
10. Kim, E. S. et al. Broadening eligibility criteria to make clinical trials more representative: American Society of Clinical Oncology and Friends of Cancer Research Joint Research Statement. *J. Clin. Oncol.* **35**, 3737–3744 (2017).
11. Labrecque, J. A. & Swanson, S. A. Target trial emulation: teaching epidemiology and beyond. *Eur. J. Epidemiol.* **32**, 473–475 (2017).
12. Danaei, G., García Rodríguez, L. A., Cantero, O. F., Logan, R. W. & Hernán, M. A. Electronic medical records can be used to emulate target trials of sustained treatment strategies. *J. Clin. Epidemiol.* **96**, 12–22 (2018).
13. Woo, M. An AI boost for clinical trials. *Nature* **573**, S100–S102 (2019).
14. Kang, T. et al. ElilE: an open-source information extraction system for clinical trial eligibility criteria. *J. Am. Med. Inform. Assoc.* **24**, 1062–1071 (2017).
15. Ni, Y. et al. Increasing the efficiency of trial–patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. *BMC Med. Inform. Decis. Mak.* **15**, 28 (2015).
16. Jonnalagadda, S. R., Adupa, A. K., Garg, R. P., Corona-Cox, J. & Shah, S. J. Text mining of the electronic health record: an information extraction approach for automated identification and subphenotyping of HFPEF patients for clinical trials. *J. Cardiovasc. Transl. Res.* **10**, 313–321 (2017).
17. Ni, Y. et al. Will they participate? Predicting patients' response to clinical trial invitations in a pediatric emergency department. *J. Am. Med. Inform. Assoc.* **23**, 671–680 (2016).
18. Miotto, R. & Weng, C. Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials. *J. Am. Med. Inform. Assoc.* **22**, e141–e150 (2015).
19. Yuan, C. et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. *J. Am. Med. Inform. Assoc.* **26**, 294–305 (2019).
20. Zhang, K. & Demner-Fushman, D. Automated classification of eligibility criteria in clinical trials to facilitate patient–trial matching for specific patient populations. *J. Am. Med. Inform. Assoc.* **24**, 781–787 (2017).
21. Shivade, C. et al. Textual inference for eligibility criteria resolution in clinical trials. *J. Biomed. Inform.* **58**, S211–S218 (2015).
22. Sen, A. et al. Correlating eligibility criteria generalizability and adverse events using big data for patients and clinical trials. *Ann. NY Acad. Sci.* **1387**, 34–43 (2017).
23. Li, Q. et al. Assessing the validity of a a priori patient–trial generalizability score using real-world data from a large clinical data research network: a colorectal cancer clinical trial case study. *AMIA Annu. Symp. Proc.* **2019**, 1101–1110 (2019).
24. Kim, J. H. et al. Towards clinical data-driven eligibility criteria optimization for interventional COVID-19 clinical trials. *J. Am. Med. Inform. Assoc.* **28**, 14–22 (2021).
25. Abernethy, A. P. et al. Real-world first-line treatment and overall survival in non-small cell lung cancer without known EGFR mutations or ALK rearrangements in US community oncology setting. *PLoS ONE* **12**, e0178420 (2017).
26. Khozin, S. et al. Real-world progression, treatment, and survival outcomes during rapid adoption of immunotherapy for advanced non-small cell lung cancer. *Cancer* **125**, 4019–4032 (2019).
27. Ma, X. et al. Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron Health, SEER, and NPCR. Preprint at https://doi.org/10.1101/2020.03.16.20037143 (2020).
28. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 4765–4774 (2017).

Methods

Clinical trial curation

In this study, we focused on aNSCLC, because aNSCLC is a prevalent cancer type and has the largest number of patients in the Flatiron Health database. We systematically identified all of the aNSCLC trials that are available for our analysis. A total of 3,684 interventional clinical trials of NSCLC were retrieved from the ClinicalTrials.gov website of the National Library of Medicine (queried on 8 November 2019). A systematic selection of trials was carried out using the following filters: (1) trials were interventional and only had two arms; (2) treatments consisted of drugs or biologicals only; (3) the drugs selected in each arm are recommended for aNSCLC as listed on the NIH website (https://www.cancer.gov/about-cancer/treatment/drugs/lung); (4) at least 250 patients in each arm were found in the Flatiron Health dataset who match the description of the patients in the trials; (5) the trial was conducted in phase III; and (6) protocols were available. The final list of selected aNSCLC trials included FLAURA[29], LUX8[30], Checkmate017[31], Checkmate057[32], Checkmate078[33], Keynote010[34], Keynote189[35], Keynote407[36], BEYOND[37] and OAK[38]. Detailed information on these trials can be found in Extended Data Table 1. To ensure the completeness of the trial criteria, we carefully extracted all of the eligibility rules directly from the original trial protocols rather than from ClinicalTrials.gov. The eligibility criteria were extracted from the original clinical trial protocol documents and the programmatic encoding of the criteria was verified by a team of experienced oncology data scientists and clinical trial specialists. Additional information about the encoding of the criteria is provided in the Supplementary Methods and Supplementary Discussion. Trial Pathfinder is a flexible framework that can be applied to other clinical trials.

Flatiron Health dataset

The data that support the findings of this study have been obtained by Flatiron Health, a nationwide EHR-derived de-identified database containing 219,312 patients with cancer with an average of 2.6 years of follow-up. The Flatiron data leveraged in this study (the February 2020 data cut) comes from a combination of EHR-derived data and external commercial and US Social Security Death Index data. The Flatiron Health database is considered one of the industry's leading research databases in oncology owing to the rigorous data curation and abstraction processes as well as publications in which their efforts to validate outcomes are demonstrated. In previous validation studies in which the Flatiron mortality data are compared to data from the gold-standard National Death Index, the sensitivity of mortality capture in a population of patients with aNSCLC was shown to be 91%, and that the effect of the remaining missing deaths on survival analyses was minimal[39,40]. In addition to curation accuracy, the Flatiron data are harmonized and aggregated across approximately 280 cancer clinics across the country, which enables its data to be more representative than the EHRs of a single healthcare centre. The majority of patients in the database originate from community oncology settings; relative community/academic proportions may vary depending on the study cohort. Data provided to investigators was de-identified and subject to obligations to prevent re-identification and to protect the confidentiality of the patients. These de-identified data may be made available upon request, and are subject to a licence agreement with Flatiron Health; interested researchers can contact DataAccess@flatiron.com to determine licensing terms. Institutional Review Board approval with a waiver of informed consent was obtained before the study was conducted.

Flatiron Health takes a comprehensive approach to data curation, which involves the collection of both structured and unstructured data from the EHRs. Structured data points, such as laboratory test results, are harmonized across different EHRs and mapped into common terminologies. Unstructured data processing, such as data that come from clinician notes or biomarker reports, leverages technology-enabled abstraction. Through this process, qualified abstractors extract key data points from unstructured documents and are aided by software that facilitates this process through organization, searching and surfacing of key documents throughout the abstraction process. Flatiron's network of abstractors includes certified tumour registrars, oncology nurses and oncology clinical researchers.

Patients in the Flatiron Health network were considered to be part of the aNSCLC real-world cohort if they were diagnosed with lung cancer (the ninth revision of the international classification of diseases (ICD-9) code 162.x; or the tenth revision of the international classification of diseases (ICD-10) code C34x or C39.9); had at least two documented clinical visits on or after 1 January 2011; had pathology consistent with NSCLC; and were diagnosed with stage IIIB, IIIC, IVA or IVB NSCLC on or after 1 January 2011, or diagnosed with early-stage NSCLC and subsequently developed recurrent or progressive disease on or after 1 January 2011. Patients were excluded if there was a lack of relevant unstructured documents in the Flatiron Health database for review by the abstraction team.

A catalogue of the criteria that it was possible to emulate using the Flatiron Health database can be found in Supplementary Table 1. There are some criteria for which Flatiron Health does not currently abstract information from EHRs—for example, reproductive health, some prior co-morbidities, some previous treatments, imaging procedures and results—and these were not included in the present study. For those criteria that are available in the database, we also evaluated the percentage of missing ECOG and laboratory value information for each patient at the start of the first or second line of therapy (Supplementary Table 38). To closely mirror the actual trial screenings, we considered clinical measurements taken within a window from 30 days before to 7 days after the start of the line of therapy[40].

Data on adverse events

We further support our findings by analysing toxicity data for a real-world cohort of 1,000 patients with aNSCLC from the Flatiron database. These patients were randomly selected from the broader aNSCLC cohort based on receipt of anti-PD-1/PD-L1 therapy, and underwent additional data abstraction to determine the reasons for treatment discontinuation, including toxicity. In addition, we identified 22 Roche oncology trials with available clinical study reports, and extracted statistics from the study reports on the number of patients who withdrew from treatment owing to adverse events.

The Trial Pathfinder workflow

In the first step of Trial Pathfinder—trial emulation—we identified individuals in the real-world dataset who met the available eligibility criteria as originally published in the clinical trial protocol. The eligibility criteria were encoded as logic statements and were automatically applied by our workflow. More information on how the semi-structured free-text criteria in the clinical trial protocols were encoded into programmatic statements is provided in the Supplementary Methods. Patients with missing data points (for example, ECOG or laboratory values) in the corresponding criteria were not filtered by those criteria. We then assigned the selected patients to the treatment groups that were consistent with their treatment records in the database (for example, atezolizumab versus docetaxel). To emulate the randomization and blind assignment in the trials, we used inverse probability of treatment weighting (IPTW) to adjust for baseline confounding factors. Time zero was set to be the start of the corresponding line of therapy. Finally, we performed survival analysis for the emulated trials using the hazard ratio of the overall survival as the outcome. Each individual was followed until the occurrence of death or censored at the latest reported activity. Outcomes that occur after 27 months in the Flatiron database are considered censored in our analysis to match the original trial settings. The results are robust to the specific window lengths discussed here (Supplementary Table 39). The Trial Pathfinder open source code was written in Python version 3.6.

Trial Pathfinder trial emulation and survival analysis

To emulate the blind assignment and obtain unbiased estimates of treatment effects, we used IPTW to adjust for the baseline covariates. During the survival analysis, patient i is given the weight defined in equation (1), in which Z_i is the indicator variable representing whether patient i is treated or not, with $Z_i = 1$ indicating a treated case. The propensity score e_i is defined in equation (2), in which X_i denotes the baseline covariates. We used a logistic regression model to estimate e_i. In our experiments of aNSCLC, the covariates X were: age, gender, composite race or ethnicity, histology, smoking status, staging, ECOG and biomarker status, including ALK, EGFR, PDL1, ROS1, KRAS and BRAF. Adjustment by propensity score is effective in balancing all of the covariates between the synthetic treatment and control groups (Extended Data Fig. 3).

$$\omega_i = Z_i/e_i + (1 - Z_i)/(1 - e_i) \quad (1)$$

$$e_i = \Pr(Z_i = 1|X_i) \quad (2)$$

We further performed survival analysis on the emulated trials. For each patient, the index date or time zero, resembling the randomization point in a clinical trial, was chosen to be the start date of the line of therapy of that trial (either first or second). This choice of time zero ensures that there is no immortal time bias[41]. Patients were followed until the occurrence of death, censoring those patients without a death event. The Cox proportional-hazards model was used to compute hazard ratios and confidence intervals of overall survival. Survival curves were estimated with the Kaplan–Meier method.

Eligibility criteria evaluation with Shapley values

To evaluate the influence of an individual criterion we used the Shapley value, which is the average expected marginal contribution of adding one criterion to the hazard ratio after all possible combinations of criteria have been considered. The Shapley value has recently been proposed in machine learning as a principled approach to quantify the contribution of individual features and data[28]. The definition of the Shapley value of the ith criterion is given in equation (3), in which n is the total number of criteria and HR(S) indicates the hazard ratio computed when the criteria subset S is used to select patients. The sum in equation (3) is taken over all possible subsets S of the n original criteria (denoted as N for short) that did not contain i.

$$\text{Shapley value of the } i\text{th criterion} = \sum_{S \subseteq N \setminus \{i\}} (|S|!\,(n - |S| - 1)!/n!)(\text{HR}(S \cup \{i\}) - \text{HR}(S)) \quad (3)$$

The Shapley value of the ith criterion is a weighted average of the effect of adding this criterion to different subsets of inclusion/exclusion criteria. The weights normalize for the number of possible sets that have the same cardinality and are required to satisfy the Shapley attribution properties.

Exhaustively computing the hazard ratios of overall survival for all possible subsets of criteria (order of $n!$) was computationally prohibitive. Here we estimated the Shapley value by Monte Carlo sampling subsets of criteria S. The Monte Carlo sampling gives an unbiased estimate of the Shapley value. Following the previously proposed algorithm[42], we stop sampling when the Shapley estimate has converged (that is, when the standard error of the Monte Carlo mean is less than 0.001). In practice, convergence happened after a hundred iterations for each criterion. A few thousand Monte Carlo samples combined is sufficient for a trial with tens of criteria to evaluate. This makes Trial Pathfinder computationally efficient (Extended Data Fig. 4) and only needs around half an hour to run with a single CPU for one trial. For each trial, we averaged its results evaluating on a different criteria set from the trials in the same line of therapy (either first or second). A Shapley value larger than zero indicates that the contribution of that criterion is to increase the hazard ratio on average. Conversely, a negative Shapley value means that the contribution of that criterion is to decrease the hazard ratio on average. Finally, Shapley values that are close to zero correspond to a criterion that does not affect the hazard ratio.

Trial Pathfinder reports the subset of criteria used by the original trial that have a Shapley value smaller than 0 as data-driven criteria. Once the data-driven subset of criteria was selected, Trial Pathfinder computed the number of eligible patients and the hazard ratio of the overall survival between the synthetic treatment and control arms.

Additional validation analyses

We stratified our 61,094 patients with aNSCLC from the Flatiron database by their geography of residence as in the US census—Northeast (n = 11,777), Midwest (n = 8,895), South (n = 23,895) and West (n = 9,061). We then evaluated the inclusion/exclusion criteria selected by Trial Pathfinder for each of the 10 aNSCLC trials for patients from each geographical region separately (Supplementary Tables 22–25). We also stratified our aNSCLC cohort by their insurance plan as an additional robustness analysis—commercial health plans (n = 22,423), Medicare (n = 10,841) and the remaining patients (n = 22,361). We evaluated our previously selected inclusion/exclusion criteria for each of the 10 aNSCLC trials for patients under the three types of insurance plans separately (Supplementary Tables 26–28). We used the nationwide (US-based) de-identified Flatiron Health-Foundation Medicine aNSCLC clinicogenomic database (FH-FMI CGDB) for further validation[43]. Genomic alterations were identified through comprehensive genomic profiling of more than 300 cancer-related genes on the next-generation sequencing-based FoundationOne panel of the FMI[44]. Retrospective longitudinal clinical data were derived from EHR data from clinics in the Flatiron network, consisting of patient-level structured and unstructured data, curated by technology-enabled abstraction, and were linked to genomic data derived from comprehensive genomic profiling tests of the FMI in the FH-FMI CGDB by de-identified and deterministic matching[43]. To leverage the rich genomics information of FH-FMI CGDB, we added 17 additional genes to the adjustment of the covariates that have alterations in at least 1,000 patients (Supplementary Table 31). For each of the 10 aNSCLC trials, we applied the inclusion/exclusion criteria that Trial Pathfinder selected on the Flatiron data and used it to emulate a trial using the FH-FMI CGDB cohort (Supplementary Table 30). Progression is used as the end point and progression-free survival hazard ratios are computed.

Statistical analysis

We bootstrapped the cohorts to estimate the standard deviations for the Shapley values. The confidence intervals for the hazard ratios were estimated from the variance matrix of the coefficients in fitting the Cox proportional-hazards model. For the safety impact analysis on 22 Roche oncology trials, we use two-sided P values from Fisher's exact tests to measure the difference in the withdrawal ratio given two sets of trials (Supplementary Table 35). When analysing toxicity data, we use two-sided P values from two-tailed Student's t-tests to evaluate whether there is a significant difference in the baseline laboratory values between two toxicity groups (Extended Data Fig. 6).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The Flatiron Health and the FH-FMI CGDB data used in this study were licensed from Flatiron Health (https://flatiron.com/real-world-evidence/) and Foundation Medicine. These de-identified

data may be made available upon request; interested researchers can contact DataAccess@flatiron.com and cgdb-fmi@flatiron.com. Information on the clinical studies can be found on clinicaltrials.gov and EUdraCT.

Code availability

The open source Python code for Trial Pathfinder is available on GitHub (https://github.com/RuishanLiu/TrialPathfinder).

29. Soria, J.-C. et al. Osimertinib in untreated EGFR-mutated advanced non-small-cell lung cancer. *N. Engl. J. Med*. **378**, 113–125 (2018).
30. Soria, J.-C. et al. Afatinib versus erlotinib as second-line treatment of patients with advanced squamous cell carcinoma of the lung (LUX-Lung 8): an open-label randomised controlled phase 3 trial. *Lancet Oncol*. **16**, 897–907 (2015).
31. Brahmer, J. et al. Nivolumab versus docetaxel in advanced squamous-cell non-small-cell lung cancer. *N. Engl. J. Med*. **373**, 123–135 (2015).
32. Borghaei, H. et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. *N. Engl. J. Med*. **373**, 1627–1639 (2015).
33. Wu, Y.-L. et al. Nivolumab versus docetaxel in a predominantly Chinese patient population with previously treated advanced NSCLC: CheckMate 078 randomized phase III clinical trial. *J. Thorac. Oncol*. **14**, 867–875 (2019).
34. Herbst, R. S. et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *Lancet* **387**, 1540–1550 (2016).
35. Gandhi, L. et al. Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer. *N. Engl. J. Med*. **378**, 2078–2092 (2018).
36. Paz-Ares, L. et al. Pembrolizumab plus chemotherapy for squamous non-small-cell lung cancer. *N. Engl. J. Med*. **379**, 2040–2051 (2018).
37. Zhou, C. et al. BEYOND: a randomized, double-blind, placebo-controlled, multicenter, phase III study of first-line carboplatin/paclitaxel plus bevacizumab or placebo in Chinese patients with advanced or recurrent nonsquamous non-small-cell lung cancer. *J. Clin. Oncol*. **33**, 2197–2204 (2015).
38. Rittmeyer, A. et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *Lancet* **389**, 255–265 (2017).
39. Curtis, M. D. et al. Development and validation of a high-quality composite real-world mortality endpoint. *Health Serv. Res*. **53**, 4460–4476 (2018).
40. Carrigan, G. et al. An evaluation of the impact of missing deaths on overall survival analyses of advanced non-small cell lung cancer patients conducted in an electronic health records database. *Pharmacoepidemiol. Drug Saf*. **28**, 572–581 (2019).
41. Suissa, S. Immortal time bias in pharmaco-epidemiology. *Am. J. Epidemiol*. **167**, 492–499 (2008).
42. Ghorbani, A. & Zou, J. Data shapley: equitable valuation of data for machine learning. In *International Conference on Machine Learning* 2242–2251 (2019).
43. Singal, G. et al. Association of patient characteristics and tumor genomics with clinical outcomes among patients with non-small cell lung cancer using a clinicogenomic database. *J. Am. Med. Assoc*. **321**, 1391–1399 (2019).
44. Frampton, G. M. et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol*. **31**, 1023–1031 (2013).

Acknowledgements We thank T. Ton, D. Hibar, A. Bier, D. Heinzmann, M. Beattie, A. Kelman, M. Heidelberg, J. Law and L. Tian for comments and discussions; M. D'Andrea, M. Lim and H. Rangi for help with the clinical trials data and M. Hwang for administrative support. S.R., S.W., N.P., A.L.P., M.L., B.A., W.C. and R.C. are supported by funding from Roche. J.Z. is supported by NSF CAREER and grants from the Chan-Zuckerberg Initiative. Y.L. is supported by 1UL1TR003142 and 4P30CA124435 from National Institutes of Health.

Author contributions R.L., A.L.P., S.R., S.W., R.C. and J.Z. designed the study. R.L., S.R., S.W. and N.P. carried out the analysis. R.L. and S.R. wrote the code. M.L., B.A., Y.L. and W.C. provided clinical interpretations. R.L., S.R., S.W., N.P., R.C. and J.Z. drafted the manuscript. R.C. and J.Z. supervised the study. All of the authors provided discussion points, and reviewed and approved the final manuscript.

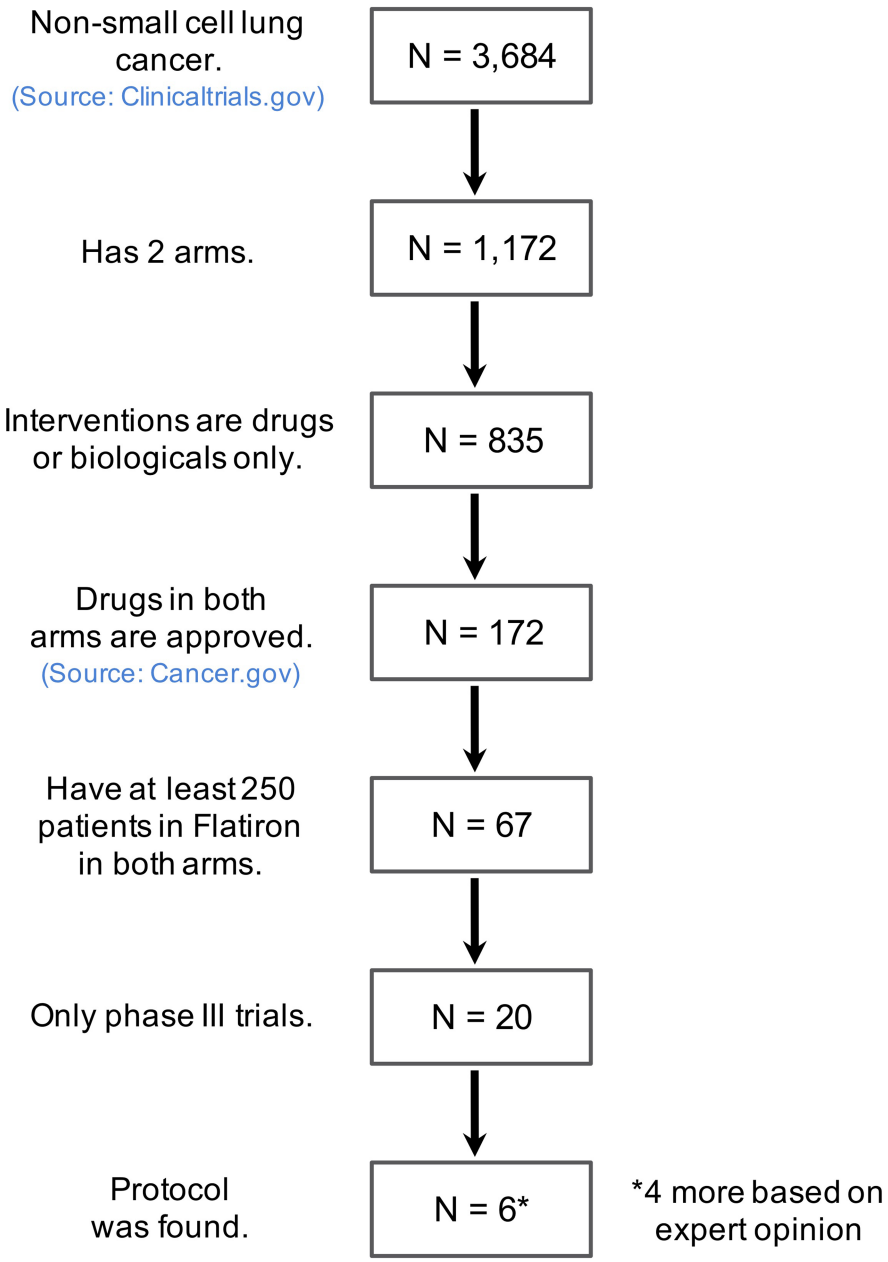
Competing interests A.L.P., S.R., M.L., B.A., N.P., S.W., R.C. and W.C. are employees of Genentech, a member of the Roche Group. R.L., Y.L. and J.Z. declare no competing interests.

Additional information

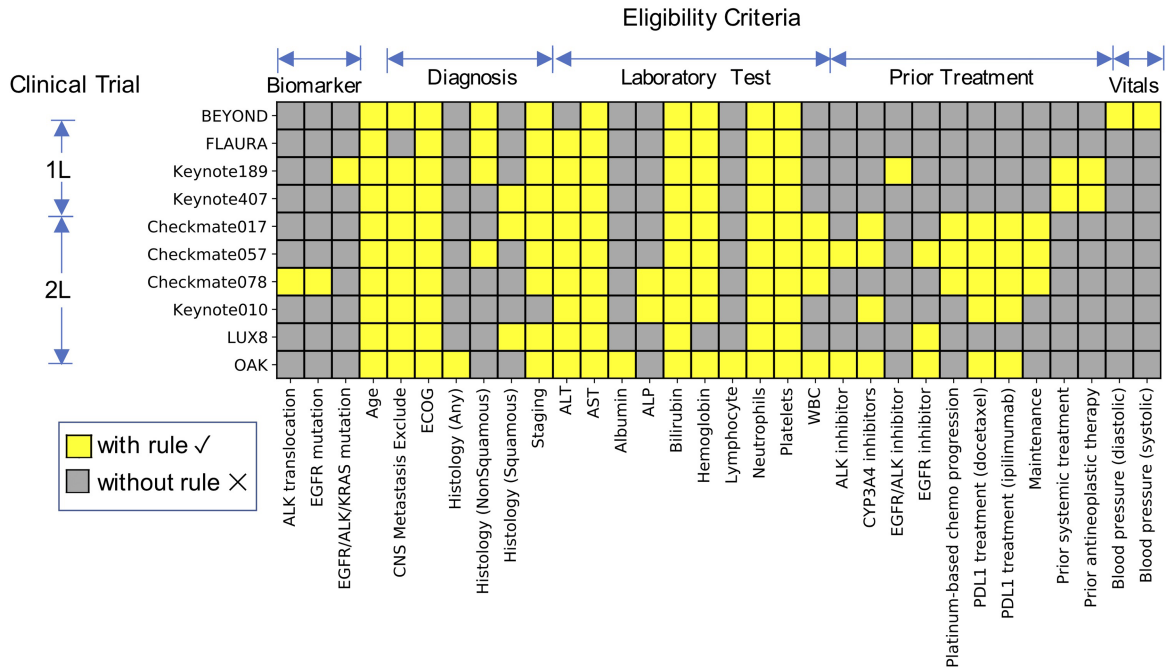
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03430-5.

Correspondence and requests for materials should be addressed to R.C. or J.Z.

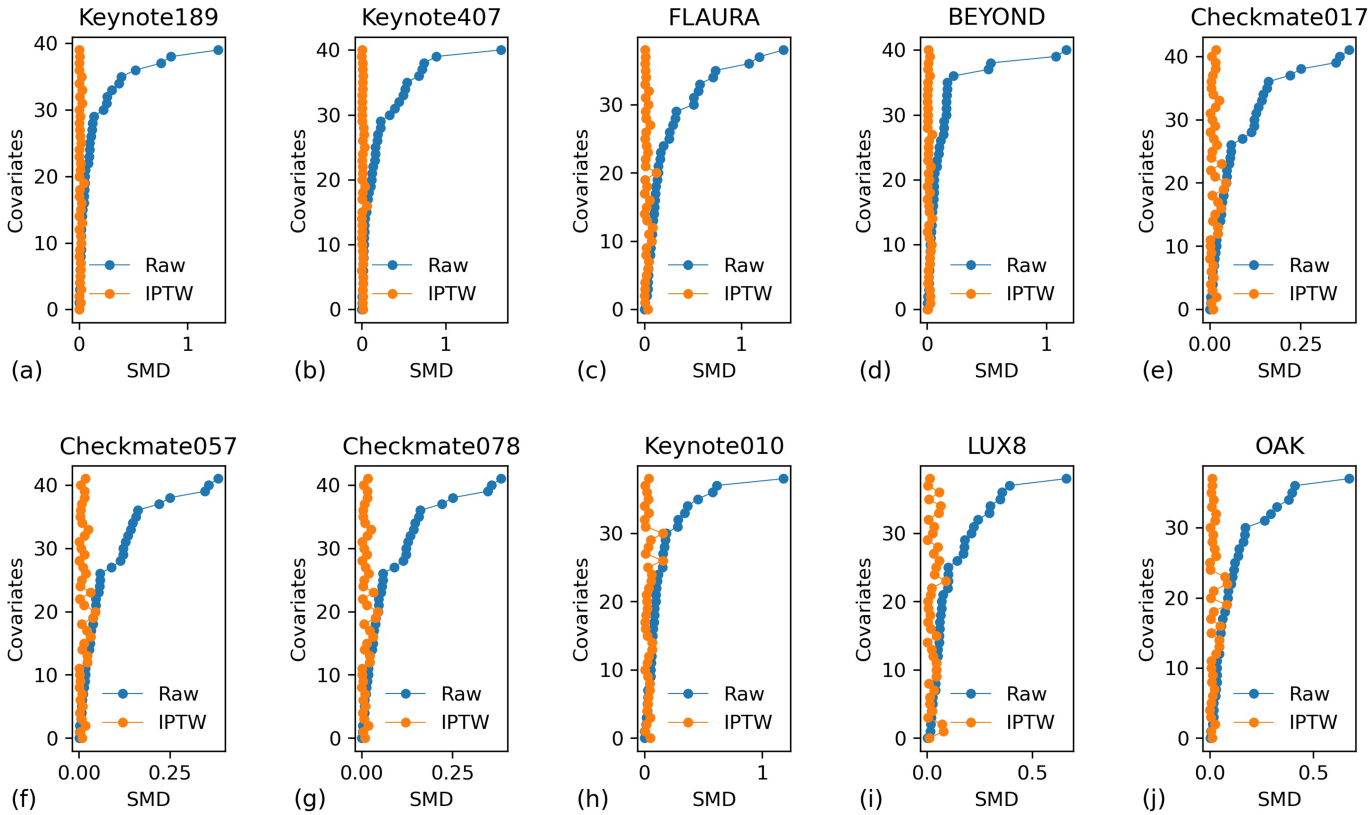
Peer review information *Nature* thanks Richard Hooper and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



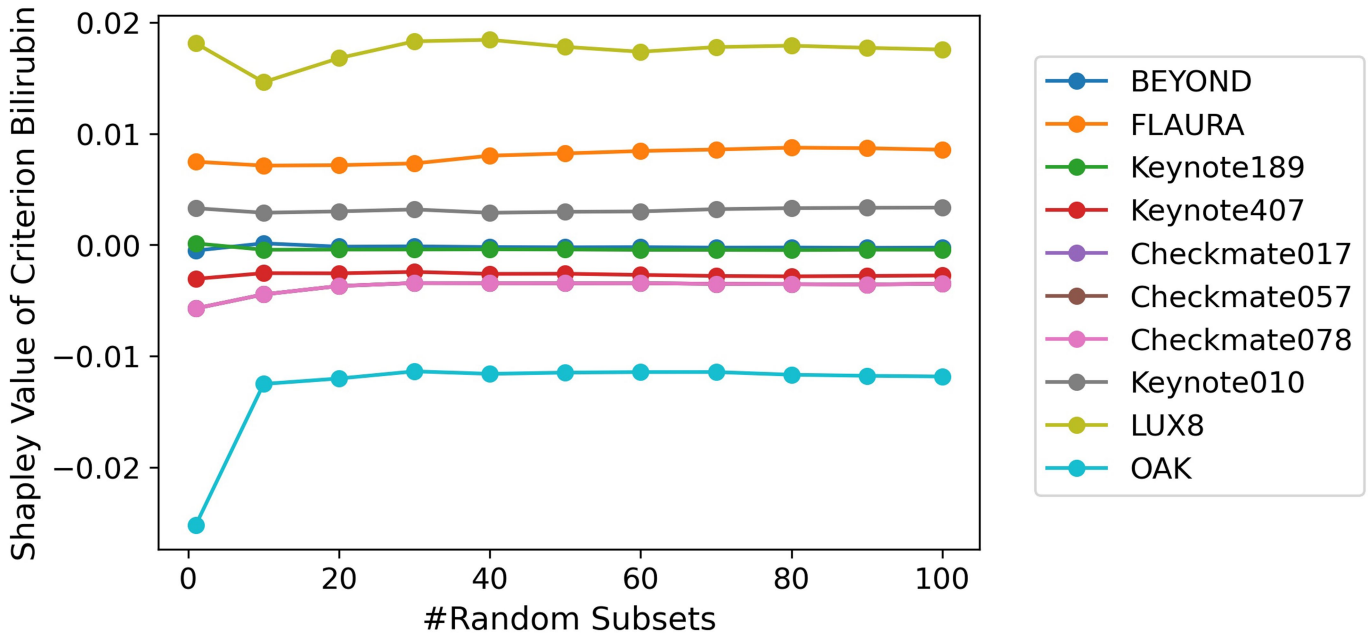
Extended Data Fig. 1 | Selection of aNSCLC clinical trials. Workflow implemented in a Python script to perform a systematic selection of trials using the six filters described in the Methods. Twenty clinical trials met the first five filters, but only six of them had a protocol that was publicly available either on ClinicalTrials.gov or as supplementary material in the associated publications. Additionally, four trials were included in the model that were suggested by subject matter experts at Roche. These four trials had not originally been identified by our systematic search owing to errors in their clinicaltrials.gov entries (for example, one trial was listed as having eight arms despite having only two).



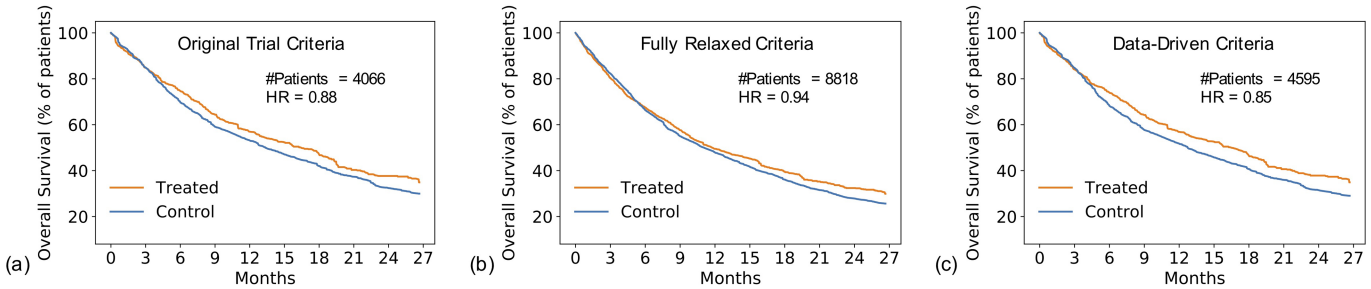
Extended Data Fig. 2 | Differential use of eligibility criteria. The trial and criteria grid shows which eligibility criteria are present in each aNSCLC trial (criteria coloured in yellow are included in the trial protocol). The trials are divided into first-line and second-line therapies, depending on their protocol design; the eligibility criteria are grouped into categories depending on the type of variable that is measured.



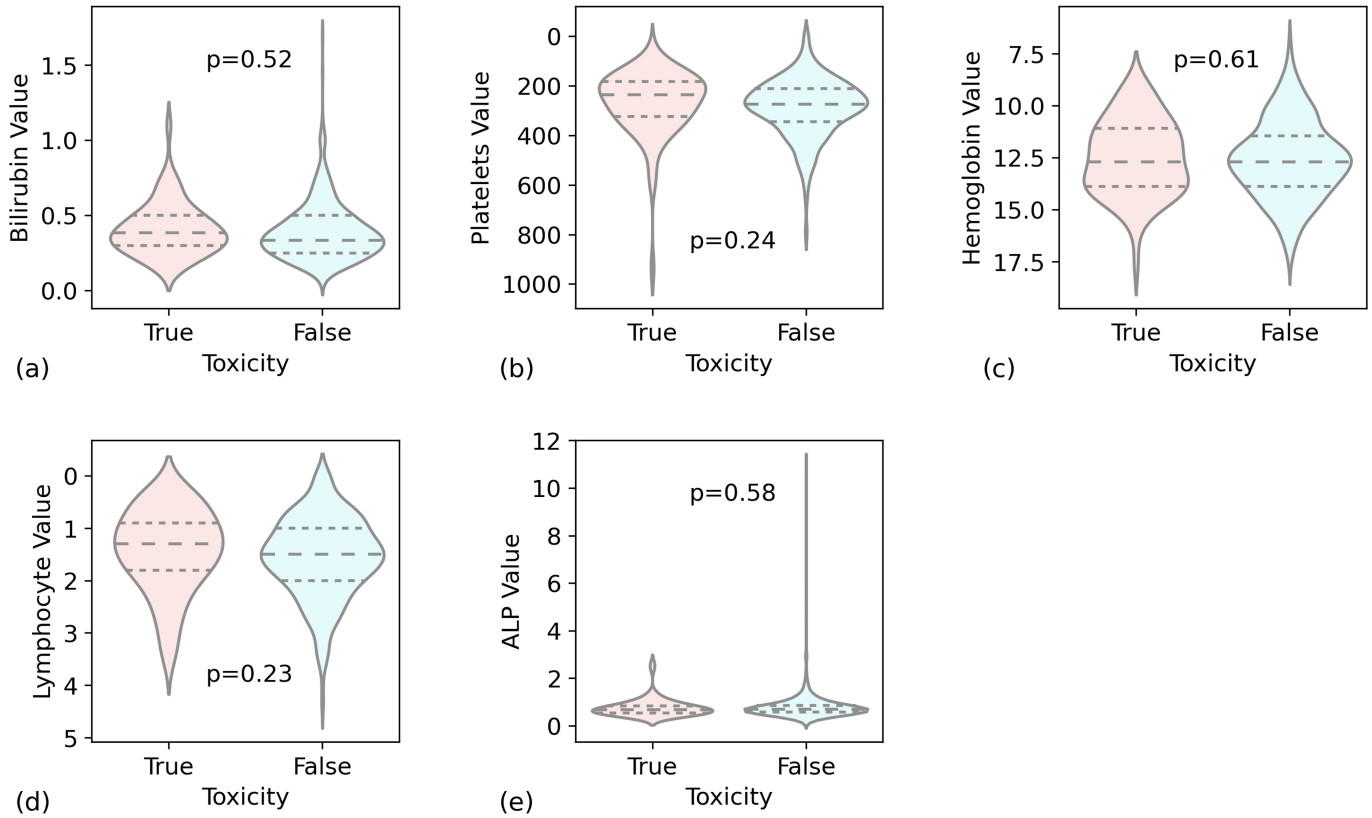
Extended Data Fig. 3 | Balance assessment for treatment and control groups. a–j, For each aNSCLC trial, we plot the standardized mean difference (SMD) for every patient covariate between the treatment and control cohorts generated from the Flatiron data. SMD values close to 0 indicate that the cohorts are balanced. The inverse propensity weighting used in our analysis (IPTW) effectively balances the cohort. 'Raw' corresponds to the unadjusted cohorts.



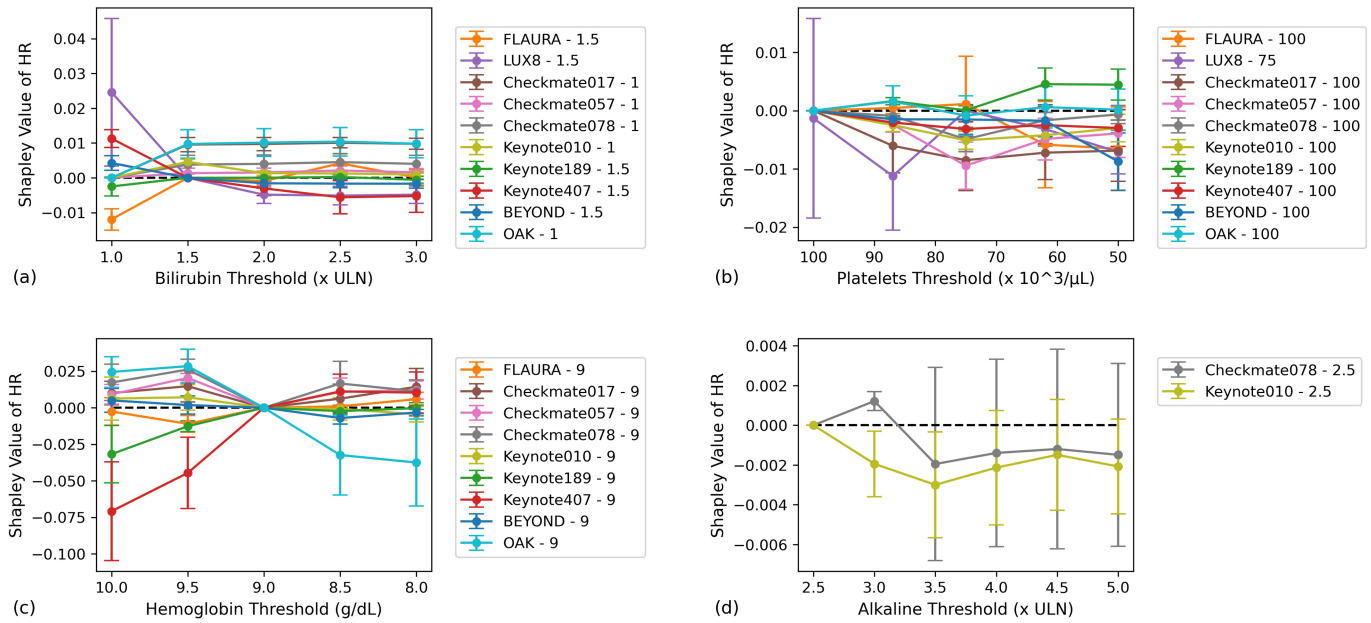
Extended Data Fig. 4 | Convergence of the Shapley value for the bilirubin criterion. The *x* axis indicates the number of randomly generated subsets of criteria used for Shapley value computation.



Extended Data Fig. 5 | Example of the effect of relaxing the eligibility criteria. a–**c**, Survival curves, hazard ratios and the number of patients in trial Keynote189 when the eligibility criteria scenarios are: the original trial criteria (**a**), fully relaxed criteria (that is, all of the patients who took the relevant treatments) (**b**) and the data-driven criteria identified by Trial Pathfinder (**c**).



Extended Data Fig. 6 | Comparison of patient baselines. a–**e**, Violin plots for the laboratory values of the patients at the start of treatment. We partition the sampled patients with aNSCLC from the Flatiron database into two groups depending on whether or not they had withdrawn from first-line aNSCLC treatments due to toxicity (82 patients with toxicity = true and 918 patients with toxicity = false). The violin plots show the distribution of each of the laboratory values at the start of the trial. There is no significant difference in the baseline laboratory values between patients who later withdrew from treatment due to toxicity and the patients who did not (unadjusted two-sided Student's *t*-test; $P > 0.2$ for all five laboratory tests).



Extended Data Fig. 7 | Effects of varying laboratory cut-off values. **a**–**d**, Changes in the Shapley value of the hazard ratios of the overall survival for different laboratory values thresholds. The *x* axis corresponds to different values of the inclusion threshold for bilirubin (serum bilirubin less than threshold for inclusion) (**a**), platelets (platelet count larger than the threshold) (**b**), haemoglobin (whole-blood haemoglobin level less than the threshold) (**c**) and ALP (ALP concentration larger than the threshold) (**d**). Changing a threshold to the right on the *x* axis corresponds to more relaxed criteria that would include more patients. The thresholds used in the original trials are provided in the key and their Shapley values are set as the baseline 0. For most of the trials, relaxing the laboratory value thresholds would not significantly change the hazard ratio or would decrease the hazard ratio (that is, curve below 0). The range of values shown for each laboratory test corresponds to the range of thresholds used in actual trials (Supplementary Table 35). In all of the panels, the error bars correspond to the bootstrap standard deviation and the centres correspond to the bootstrap mean of five replications.

Extended Data Table 1 | Summary of investigated aNSCLC trials

Trial Short Code	Sponsor	ClinicalTrials ID	LoT	Experiment	Control	Published HR (95% CI)
FLAURA	AstraZeneca	NCT02296125	1L	osimertinib	erlotinib or gefitinib	0.63 (0.45, 0.88)
LUX8	Boehringer Ingelheim	NCT01523587	2L	afatinib	erlotinib	0.81 (0.69, 0.95)
Checkmate017	Bristol Myers Squibb	NCT01642004	2L	nivolumab	docetaxel	0.59 (0.44, 0.79)
Checkmate057	Bristol Myers Squibb	NCT01673867	2L	nivolumab	docetaxel	0.73 (0.59, 0.89)
Checkmate078	Bristol Myers Squibb	NCT02613507	2L	nivolumab	docetaxel	0.68 (0.52, 0.90)
Keynote010	Merck	NCT01905657	2L	pembrolizumab	docetaxel	0.71 (0.58, 0.88) 0.61 (0.49, 0.75)
Keynote189	Merck	NCT02578680	1L	pembrolizumab plus carboplatin or cisplatin, and pemetrexed	Placebo plus carboplatin or cisplatin, and pemetrexed	0.49 (0.38, 0.64)
Keynote407	Merck	NCT02775435	1L	pembrolizumab + carboplatin + (paclitaxel OR nab-paclitaxel)	Placebo + carboplatin + (paclitaxel OR nab-paclitaxel)	0.64 (0.49, 0.85)
BEYOND	Roche	NCT01364012	1L	bevacizumab + carboplatin + paclitaxel	Placebo + Carboplatin + paclitaxel	0.68 (0.50, 0.93)
OAK	Roche	NCT02008227	2L	atezolizumab	docetaxel	0.73 (0.62, 0.87)

Line of therapy (LoT); confidence interval (CI); first line of therapy (1L); second line of therapy (2L).

Extended Data Table 2 | Validation on progression-free survival hazard ratio

Trial Name	Original Trial Criteria			Fully Relaxed Criteria		Data-driven Criteria Learned from OS Hazard Ratio		
	#Criteria	#Patients	HR (%95 CI)	#Patients	HR (%95 CI)	#Criteria	#Patients	HR (%95 CI)
FLAURA	10	2277	0.64 (0.52, 0.79)	3819	0.65 (0.55, 0.77)	4	2546	0.61 (0.50, 0.74)
LUX8	11	129	0.80 (0.49, 1.30)	1350	1.13 (0.98, 1.29)	5	141	0.66 (0.45, 0.97)
Checkmate 017	17	523	0.59 (0.45, 0.78)	4900	0.77 (0.72, 0.83)	7	4085	0.77 (0.71, 0.84)
Checkmate 057	19	792	0.88 (0.74, 1.04)	4900	0.77 (0.72, 0.83)	9	2594	0.77 (0.70, 0.85)
Checkmate 078	18	1509	0.75 (0.66, 0.86)	4900	0.77 (0.72, 0.83)	9	3348	0.74 (0.67, 0.81)
Keynote010	13	806	0.60 (0.49, 0.73)	1950	0.54 (0.48, 0.62)	1	1948	0.55 (0.48, 0.62)
Keynote189	15	4066	0.81 (0.71, 0.93)	8818	0.82 (0.76, 0.89)	7	4595	0.81 (0.72, 0.91)
Keynote407	13	2031	1.22 (0.96, 1.54)	10437	1.14 (1.04, 1.25)	4	9173	1.14 (1.03, 1.26)
BEYOND	12	2902	1.09 (1.00, 1.19)	9310	1.13 (1.04, 1.23)	4	3043	1.09 (1.01, 1.19)
OAK	19	493	1.07 (0.82, 1.41)	1288	0.89 (0.76, 1.04)	6	620	0.98 (0.77, 1.24)
Average	15	1553	0.85	5167	0.86	6	3209	0.81

The number of inclusion and exclusion criteria, the number of eligible patients and the hazard ratio of progression-free survival with confidence interval of emulated aNSCLC trials with eligibility criteria under three scenarios: original criteria of the clinical trial, fully relaxed criteria and data-driven criteria learned from results of the hazard ratio of overall survival (same as in Table 1).

Extended Data Table 3 | Analysis in other cancers

Cancer Type	Trial Name	Original Trial Criteria			Fully Relaxed Criteria		Data-driven Criteria		
		#Criteria	#Patients	HR	#Patients	HR	#Criteria	#Patients	HR
CRC	PRIME	11	1742	0.67 (0.49, 0.92)	3048	0.67 (0.53, 0.84)	4	2680	0.64 (0.50, 0.82)
Advanced Melanoma	COMBIv	14	393	0.86 (0.65, 1.14)	720	0.88 (0.72, 1.08)	7	556	0.79 (0.63, 1.00)
Metastatic Breast	Marianne	12	101	1.55 (0.77, 3.11)	300	1.36 (0.96, 1.93)	5	182	1.26 (0.79, 2.01)
Average		12	745	1.03	1356	0.97	5	1139	0.90

Eligibility criteria for colorectal cancer (CRC), advanced melanoma and metastatic breast cancer in three scenarios. The number of inclusion and exclusion criteria, the number of eligible patients and the hazard ratio of the overall survival with confidence interval of emulated aNSCLC trials with eligibility criteria under three scenarios: original criteria of the clinical trial, fully relaxed criteria and data-driven criteria.

nature research

Corresponding author(s): James Zou, Ryan Copping

Last updated by author(s): February 28, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We did not collect our own data for this study. This study utilized the Flatiron Health electronic health record (EHR)-derived database (datacut Feb 2020), which we describe in more detail in the Data section below.
Data analysis	Data analysis was performed using custom Python code (Python version 3.6) developed by the study authors. The code is open source and is available at https://github.com/RuishanLiu/TrialPathfinder.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The Flatiron Health and the FH-FMI CGDB data used in this study were licensed from Flatiron Health, Inc. (https://flatiron.com/real-world-evidence/) and Foundation Medicine, Inc. These de-identified data may be made available upon request; interested researchers can contact DataAccess@flatiron.com and cgdb-fmi@flatiron.com. Information on the clinical studies can be found on clinicaltrials.gov and EUdraCT.

The Flatiron dataset includes de-identified data from approximately 280 cancer clinics (~800 sites of care) representing more than 2.4 million cancer patients in the

United States. Data provided to investigators was de-identified by Flatiron Health and provisions were in place to prevent re-identification in order to protect patients' confidentiality. The deidentified patient-level data in the EHRs include structured data (e.g. laboratory values, prescribed drugs) in addition to unstructured data collected via technology-enabled chart abstraction from physicians' notes and other unstructured documents (e.g. biomarker reports). We also performed validation analysis using the Flatiron Health-Foundation Medicine Inc. Clinico-Genomic Database (FH-FMI CGDB). This de-identified database consists of research-grade specimen, genomic, and other biomarker data arising from clinical testing performed by Foundation Medicine. Genomic alterations were identified via comprehensive genomic profiling of over 300 cancer-related genes on FMI's next-generation sequencing (NGS) based FoundationOne® panel. Retrospective longitudinal clinical data were derived from EHR data from clinics in the Flatiron network, consisting of patient-level structured and unstructured data, curated via technology-enabled abstraction. Institutional Review Board approval with waiver of informed consent was obtained prior to study conduct.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We focused on analyzing aNSCLC trials because they have the largest number of patients in the Flatiron dataset with 61,094 aNSCLC patients.
Data exclusions	Patients in the Flatiron Health network were considered to be part of the aNSCLC enhanced data mart if they were diagnosed with lung cancer (ICD-9: 162.x, or ICD-10: C34x or C39.9); had at least two documented clinical visits on or after January 1, 2011; had pathology consistent with non-small cell lung cancer (NSCLC); and were diagnosed with Stage IIIB, IIIC, IVA or IVB NSCLC on or after 1/1/2011, or diagnosed with early-stage NSCLC and subsequently developed recurrent or progressive disease on or after 1/1/2011. Patients were excluded if there was a lack of relevant unstructured documents in the Flatiron Health database for review by the abstraction team.
Replication	We performed 4 sets of analyses to replicate or to support the robustness of our results First, in addition to using overall survival as the end point, we repeated all of the analyses for each trial using progression-free survival (PS). The results are highly consistent. Next, we stratified the patients by their geographic region (Northeast, Midwest, West, South U.S.) and their insurance types (commercial health plans, Medicare, other coverage). The results of the analyses on each subgroup of patients are consistent with our primary findings. With the advent of immunotherapies in the cancer treatment landscape over the past few years, the standard of care has rapidly evolved for NSCLC patients. In order to assess the robustness of our findings in light of this, we also ran a sensitivity analysis in which the selected data-driven criteria was applied to patients who received treatment in the last three years (Feb 2017 to Feb 2020). The results for the recent patients are consistent with the results for the full cohort, further supporting the robustness of the Trial Pathfinder findings across time and shifts in treatment patterns. Finally, our primary analyses focused on aNSCLC trials because that is where we have the largest number of patients in Flatiron. To demonstrate that our framework and findings can be extended to other cancer types, we identified three more trials in colorectal cancer (CRC), melanoma and breast cancer with available trial protocols that can be encoded in Flatiron. Our findings in these other cancer types are consistent with the findings in aNSCLC.
Randomization	This is a retrospective analysis of real-world data. Because the data is observational, it was not possible to perform randomized interventions. Instead we followed the best practice of working with observational data and used propensity score weighting to adjust for differences between different patient cohorts.
Blinding	Our study utilized observational data derived from the EHR. Because this is real-world data, it's not feasible to blind the research participants to the treatment that they were given. Therefore blinding was not applied.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☐	☒ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	We used real-world EHR data licensed from Flatiron Health and did not conduct our own clinical trials.
Study protocol	Starting from all of the phase 3 aNSCLC trials on ClinicalTrials.gov (queried on November 8th, 2019), we filtered for trials that have available trial protocol and at least 250 patients in each arm were found in the Flatiron Health dataset that match the description of the patients in the trials. This resulted in 10 completed aNSCLC trials which we analyzed using the Trial Pathfinder framework. The protocol for each trial was obtained from the original trial publication.
Data collection	This retrospective study utilized the Flatiron Health electronic health record (EHR)-derived database (the February 2020 datacut), which includes de-identified data from over 280 cancer clinics (~800 sites of care) in the United States. Longitudinal patient-level data include structured and unstructured data curated from the EHR. Data provided to investigators was de-identified by Flatiron Health and provisions were in place to prevent re-identification in order to protect patients' confidentiality. We also performed validation analysis using the Foundation Medicine Inc. Clinico-Genomic Database (FMI CGDB). This de-identified database consists of research-grade specimen, genomic, and other biomarker data arising from clinical testing performed by Foundation Medicine. Genomic alterations were identified via comprehensive genomic profiling of over 300 cancer-related genes on FMI's next-generation sequencing (NGS) based FoundationOne® panel. Retrospective longitudinal clinical data were derived from EHR data from clinics in the Flatiron network, consisting of patient-level structured and unstructured data, curated via technology-enabled abstraction. Institutional Review Board approval with waiver of informed consent was obtained prior to study conduct.
Outcomes	We used overall survival as the primary outcome. Each patient was followed until the occurrence of death. Patients without a death event were censored at the latest structured activity.