Clinical trial design and dissemination: comprehensive analysis of clinicaltrials.gov and PubMed data since 2005

Magdalena Zwierzyna,[1,2] Mark Davies,[1] Aroon D Hingorani,[2,3] Jackie Hunter[1,4]

[1]BenevolentBio Ltd, London NW1 1LW, UK
[2]Institute of Cardiovascular Science, University College London, London, UK
[3]Farr Institute of Health Informatics, London, UK
[4]St George's Hospital Medical School, London, UK

Correspondence to:
M Zwierzyna
magda.zwierzyna@benevolent.ai
(or @magda_zw on Twitter)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2018;361:k2130
http://dx.doi.org/10.1136/bmj.k2130

Accepted: 19 April 2018

Abstract

Objective

To investigate the distribution, design characteristics, and dissemination of clinical trials by funding organisation and medical specialty.

Design

Cross sectional descriptive analysis.

Data sources

Trial protocol information from clinicaltrials.gov, metadata of journal articles in which trial results were published (PubMed), and quality metrics of associated journals from SCImago Journal and Country Rank database.

Selection criteria

All 45 620 clinical trials evaluating small molecule therapeutics, biological drugs, adjuvants, and vaccines, completed after January 2006 and before July 2015, including randomised controlled trials and non-randomised studies across all clinical phases.

Results

Industry was more likely than non-profit funders to fund large international randomised controlled trials, although methodological differences have been decreasing with time. Among 27 835 completed efficacy trials (phase II-IV), 15 084 (54.2%) had disclosed their findings publicly. Industry was more likely than non-profit trial funders to disseminate trial results (59.3% (10 444/17 627) *v* 45.3% (4555/10 066)), and large drug companies had higher disclosure rates than small ones (66.7% (7681/11 508) *v* 45.2% (2763/6119)). Trials funded by the National Institutes of Health (NIH) were disseminated more often than those of other non-profit institutions (60.0% (1451/2417) *v* 40.6% (3104/7649)). Results of studies funded by large drug companies and NIH were more likely to appear on clinicaltrials.gov than were those from non-profit funders, which were published mainly as journal articles. Trials reporting the use of randomisation were more likely than non-randomised studies to be published in a journal article (6895/19 711 (34.9%) *v* 1408/7748 (18.2%)), and journal publication rates varied across disease areas, ranging from 42% for autoimmune diseases to 20% for oncology.

Conclusions

Trial design and dissemination of results vary substantially depending on the type and size of funding institution as well as the disease area under study.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Suboptimal study design and slow dissemination of findings are common among registered clinical trials, with almost half of all studies remaining unpublished years after completion

A wider range of organisations is now funding clinical trials

Reporting rates and other quality measures may vary by the type of funding organisation

WHAT THIS STUDY ADDS

Design characteristics (as a marker of methodological quality) vary by funder and medical specialty and associate with reporting rates and the likelihood of publication in high impact journals

Results of industry funded trials are more likely to be disclosed than those from other funders, although small drug companies lag behind "big pharma"

Big pharma and NIH funded studies show high rates of reporting in clinicaltrials.gov, and trials funded by other academic and non-profit organisations were disclosed primarily as journal articles

Result dissemination rate varies by medical specialty, mainly owing to differential journal publication rates, ranging from 42% for autoimmune diseases to 20% for oncology

Introduction

Well conducted clinical trials are widely regarded as the best source of evidence on the efficacy and safety of medical interventions.[1] Trials of first in class drugs also provide the most rigorous test of causal mechanisms in human disease.[2] Findings of clinical trials inform regulatory approvals of novel drugs, key clinical practice decisions, and guidelines and fuel the progress of translational medicine.[3 4] However, this model relies on trial activity being high quality, transparent, and discoverable.[5] Randomisation, blinding, adequate power, and a clinically relevant patient population are among the hallmarks of high quality trials.[1 5] In addition, timely reporting of findings, including negative and inconclusive results, ensures fulfilment of ethical obligations to trial volunteers and avoids unnecessary duplication of research or creation of biases in the clinical knowledge base.[6 7]

To improve the visibility and discoverability of clinical trials, the US Food and Drug Administration (FDA) mandated the registration of interventional efficacy trials and public disclosure of their results.[8-11] Seventeen registries store records of clinical studies.[12 13] Clinicaltrials.gov, the oldest and largest of such platforms,[14 15] serves as both a continually updated trial registry and a database of trial results, thus offering an opportunity to explore, examine, and monitor the clinical research landscape.[10 14]

Previous studies of clinical research activity used registered trial protocols and associated published reports to investigate trends in quality, reporting rate, and potential for publication bias.[14 16-21] However, many previous studies relied on time consuming manual literature searches, which confined their scope

to pivotal randomised controlled trials or research funded by major trial sponsors. Large scale analysis of associations between a wide range of protocol characteristics and result dissemination rates, and the differences between large and small organisations, has not been previously undertaken. This is important, given a reported growth in trials funded by small industrial and academic institutions.[22-26]

To fill this gap, we did the most comprehensive analysis to date of all interventional studies of small molecule and biological therapies registered on clinicaltrials.gov, up to 19 July 2017. We linked trial registry entries to the publication of findings in journal articles and used descriptive statistics, semantic enrichment, and data visualisation methods to investigate and illustrate the landscape and evolution of registered clinical trials. Specifically, we investigated the influence of type of funder and disease area on the design and publication rates of trials across a wide range of medical journals; changes over time; and associations with reported attrition rates in drug development. To facilitate comparison across diverse study characteristics, we designed a novel visualisation method allowing rapid, intuitive inspection of basic trial properties.

Methods

Our large scale, cross sectional analysis and field synopsis of the current and historical clinical trial landscape used a combination of basic information on trial protocols and published trial results. The developed data processing and integration workflow manipulates and links data from clinicaltrials.gov with information about journal articles in which trial results were published.

Clinicaltrials.gov data processing and annotation

We downloaded the clinicaltrials.gov database as of 19 July 2017—nearly two decades after the FDA Modernization Act legislation requiring trial registration.[8] The primary dataset encompassed 249 904 studies. We then identified records of all pharmaceutical and biopharmaceutical trials, defined as clinical studies evaluating small molecule drugs, biological drugs, adjuvants, and vaccines.[27] To extract the relevant subset, we restricted the "study type" field to "interventional" and the "intervention type" to either "drug" or "biological." This led to a set of 119 840 clinical trials with a range of organisations, disease areas, and design characteristics (supplementary file 1).

To enable robust aggregate analyses, we processed and further annotated the dataset. Firstly, we harmonised date formats and age eligibility fields. Next, wherever possible, we inferred missing categories from available fields as described by Califf and colleagues.[16] For example, for single arm interventional trials with missing allocation and blinding type annotations, allocation was assigned as "non-randomised" and blinding as "open label."[16] The dataset was further enriched with derived variables. For example, we derived the trial duration from available trial start and primary completion dates, mapped trial locations to continents and flagged them if they involved emerging markets, and labelled trials making use of placebo and/or comparator controls on the basis of information from intervention name, arm type, trial title, and description field. For instance, if a study had an arm of type "placebo comparator" or an intervention with a name including words "placebo," "vehicle," or "sugar pill," it was annotated as "placebo controlled"; similarly, studies with an arm of type "active comparator" were classified as "comparator controlled" (see supplementary file 2 for details). Finally, we annotated each trial with the number of outcome measures and eligibility criteria following parsing of the unstructured eligibility data field as described by Chondrogiannis et al.[28]

To categorise clinical trials by medical specialty, we mapped reported names of disease conditions to corresponding medical subject headings (MeSH) terms (from both MeSH Thesaurus and MeSH Supplementary Concept Records)[29] by using a dictionary based named entity recognition approach.[30] A total of 21 884 distinct condition terms were annotated with MeSH identifiers, accounting for 86.4% of all disease names from 88% of the clinical trials (supplementary file 2). The mapping enabled normalisation of various synonyms to preferred disease names, as well as automatic classification of diseases into distinct categories using the MeSH hierarchy. To compare our results with previously published clinical trial success rates, we focused on seven major disease areas investigated in the study by Hay et al[31]: oncology, neurology, and autoimmune, endocrine, respiratory, cardiovascular, and infectious diseases. We mapped trial conditions annotated with MeSH terms to these categories on the basis of matching terms from levels 2 and 3 of the MeSH term hierarchy; see supplementary file 2 for details on MeSH term assignment and disease category mapping.

To categorise studies by type of funder, we first extended the agency classification available from clinicaltrials.gov. Specifically, we used two main categories (industry and non-profit organisations) and four sub-categories (big pharma, small pharma, National Institutes of Health (NIH), and other); the "other" category corresponded to universities, hospitals, and other non-profit research organisations.[32 33] We classified industrial organisations as "big pharma" or "small pharma" on the basis of their sales in 2016 as well as the overall number of registered trials they sponsored. Specifically, we classed an industrial organisation as big pharma if it featured on the list of 50 companies from the 2016 Pharmaceutical Executive ranking or if it sponsored more than 100 clinical trials[34]; otherwise, we classified it as small pharma (see supplementary file 2). Next, we derived the likely source of funding for each trial based on the "lead_sponsor" and "collaborators" data fields, using a previously described method extended to include the distinction between big and small commercial organisations.[16] Briefly, if NIH was listed either as the lead sponsor of a trial or a collaborator,

doi: 10.1136/bmj.k2130 | *BMJ* 2018;361:k2130 | the bmj

we classified the study as NIH funded. If NIH was not involved, but either the lead sponsor or a collaborator was from industry, we classified the study as funded by either big pharma if it included any big pharma organisation or otherwise as small pharma. Remaining studies were assigned "other" as the funding source.

Establishing links between trial registry records and published results

We linked the trial registry records to the disclosed results, either deposited directly on clinicaltrials.gov or published as a journal article, to establish whether the results of a particular trial had been disseminated. We implemented the approach described by Powell-Smith and Goldacre to identify eligible trials that should have publicly disclosed results.[20] These were studies concluded after January 2006 and by July 2015 (to allow sufficient time for publication of results of studies that had not filed an application to delay submission of the results (based on the "first_received_results_disposition_date" field in the registry).[20] We included studies from all organisations (even minor institutions) regardless of how many trials they funded. The selection process led to a dataset of 45 620 studies, including 27 835 efficacy trials (phase II-IV), summarised by the flow diagram in supplementary file 1.

For each eligible trial, we used two methods to search for published results. Firstly, we searched for structured reports submitted directly to the results database of clinicaltrials.gov (based on the "clinical_results" field). Next, we queried the PubMed database with NCT numbers of eligible trials, using an approach that involves searching for the NCT number in the title, abstract, and "secondary source ID" field of PubMed indexed articles.[20 35 36] To exclude study protocols, commentaries, and other non-relevant publication types, we used the broad "therapy" filter—a standard validated PubMed search filter for clinical queries[37]—as well as simple keyword matching for "study protocol" in publication titles.[20] In addition, we excluded articles published before trial completion.

We then determined the proportion of publicly disseminated trial reports for the studies covered by mandate for submission of results (efficacy trials: phase II-IV), as well as for all completed drug trials including those not covered by the FDA Amendment Act (phase 0-I trials and studies with phase set to "n/a").[9] In addition to result dissemination rate, we calculated the dissemination lag for individual studies to determine the proportion of reports disclosed within the required 12 months after completion of a trial. We defined dissemination lag as time between study completion and first public disclosure of the results (whether through submission to clinicaltrials.gov or publication in a journal).

Categorising scientific journals

To investigate the relation between characteristics of published trials and quality metrics of scientific journals, we classified journals as high or low impact based on journal metrics information from SCImago Journal and Country Rank, a publicly available resource that includes journal quality indicators developed from the information contained in the Scopus database (www.scimagojr.com). Specifically, we classified a journal as high impact if its 2016 H-index was larger than 400 and as low impact if its H-index was below 50.

Statistical analysis and data visualisation

We examined the study protocol data for 15 characteristics available from clinicaltrials.gov and from our derived annotations. These were randomisation, interventional study design (for example, parallel assignment, sequential assignment, or crossover studies), masking (open label, single blinded, double blinded studies), enrolment (total number of participants in completed studies or estimated number of participants for ongoing studies), arm types (placebo or active comparator), duration, number of outcomes (clinical endpoints), number of eligibility criteria, study phase, funding source, disease category, countries, number of locations, publication of results, and reported use of data monitoring committees. Further explanation and definitions of trial properties are available in supplementary file 2.

On the basis of these criteria characteristics, we examined the methodological quality of trials funded by different types of organisations and compared studies published in higher and lower impact journals and reported on clinicaltrials.gov. In addition, we analysed trial dissemination rates across funding categories and disease areas and compared them with clinical success rates reported previously.[31]

We used descriptive statistics and data visualisation methods to characterise trial categories. We report frequencies and percentages for categorical data and use medians and interquartile ranges for continuous variables. To visualise multiple features of clinical trial design, we used radar plots with each axis showing the fraction of trials with a different property, such as reported use of randomisation, blinding, or active comparators. We converted three continuous variables to binary categories: we classified trials as "small" if they enrolled fewer than 100 participants, as "short" if they lasted less than 365 days, and as "multi-country" if trial recruitment centres were in more than one country. Further details are available in supplementary tables. To facilitate comparisons between several radar plots, we always standardised the order and length of axes. We used Python for all data analysis tasks (scikit-learn, pandas, scipy, and numpy libraries), working with geographical data (geonamescache, geopy, and Basemap libraries), as well as for data processing and visualisation (matplotlib, Basemap, and seaborn libraries).

Patient involvement

Patients were not involved in any aspect of the study design or conduct or in the development of the research question. The study is an analysis of existing

publicly available data, so there was no active patient recruitment for data collection.

Results

Overview of clinical trial activity by funder type

To date, 119 840 drug trials have been registered with clinicaltrials.gov. As illustrated by figure 1 (left), the distribution of trials funded by different types of organisations has changed over time. In particular, the proportion of studies funded by organisations other than industry or NIH has increased—mainly universities, hospitals, and other academic and non-profit agencies, classified here as "other." Overall, these institutions funded 36% (43 431) of all registered pharmaceutical trials, followed by big pharma with 31% (36 912), and small pharma with 21% (25 216) of studies. NIH funded 11% (13 426) of registered pharmaceutical trials (fig 1, right).

Among the 10 main study funders ranked by the overall number of trials, two belonged to the NIH category (National Cancer Institute leading with 7219 trials) and eight to the big pharma category (GlaxoSmithKline at the top with 3171 trials). Of 4914 small pharma organisations, 3615 (73.6%) had fewer than five trials, and 2119 (43.1%) funded only one study. Similarly, among the 6468 funders classified as "other," 5124 (79.2%) had fewer than five trials, and 3630 (56.1%) funded only one study.

The approach outlined in the Methods section identified 45 620 clinical trials in our dataset that should also have available published results. The remaining analysis in the Results section focuses on this subset of studies.

Trial design characteristics

The dataset included various types of study designs from large randomised clinical trials to small single site studies, many of which lacked controls (supplementary file 2). Comparative analysis showed that methodological characteristics of trials varied substantially across clinical phases (fig 2, left), disease areas (fig 2, right), and funding source (fig 3). The next section will focus on a comparative analysis by funder type, and a detailed overview of differences across phases and diseases is available in supplementary file 2.

Industrial organisations were more likely to fund large international trials and more often reported the use of randomisation, controls, and blinding, compared with non-industrial funders. Although differences were evident across all clinical phases (supplementary file 2), the largest variability was observed for phase II trials, illustrated by figure 3. For instance, 4956 (68.5%) of all 7236 phase II trials funded by industry were randomised (compared with 39.4% (661/1677) of studies funded by NIH and 50.6% (1283/2535) funded by other institutions). Studies with industrial funding were also substantially larger (median 80 patients compared with 43 and 48, respectively) and more likely to include international locations (31.3% (n=2265) for industry versus 11.0% (184) for NIH and 4.0% (101) for others). Despite larger enrolment size, industry funded studies were on average shorter than trials by other funders (median 547 days compared with 1431 days for NIH and 941 for others). Non-industrial funders were more likely to report the use of data monitoring committees: 702 (41.9%) of trials with NIH funding and 1219 (48.1%) of studies funded by other organisations involved data monitoring committees compared with 1935 (26.7%) of trials funded by industry. Design characteristics did not differ substantially between trials funded by big and small pharma (supplementary file 2).

Trial design changed substantially over time, with the trends most evident in phase II. As shown in figure 4, the methodological differences between industry and non-industrial funders have generally decreased over time. Although a higher proportion of industry led studies reported the use of randomisation overall, the proportion of randomised clinical trials funded

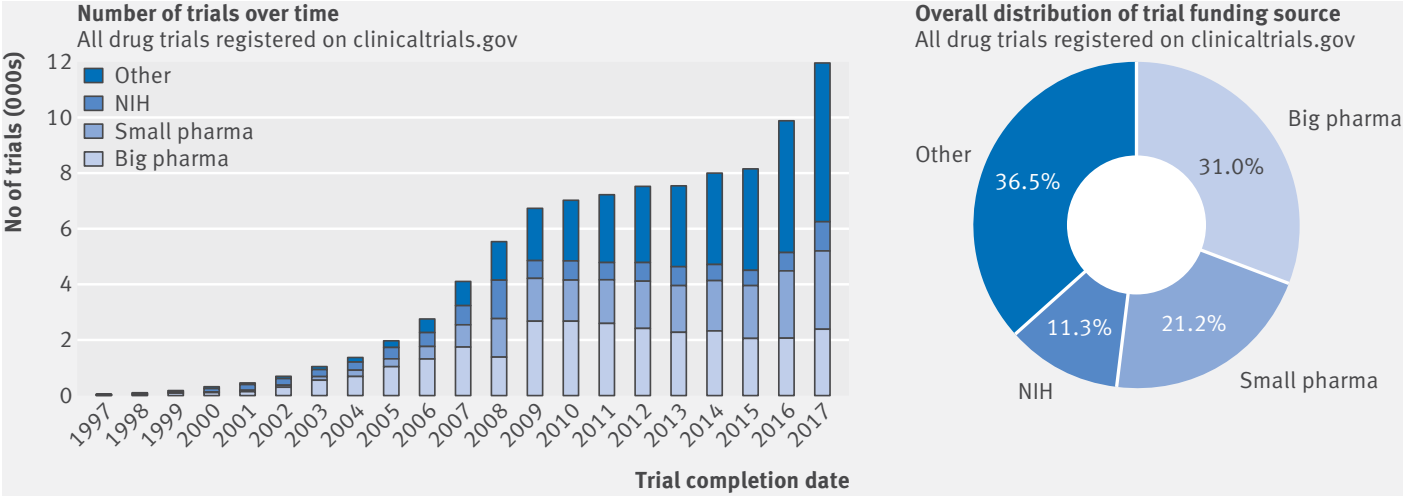

Fig 1 | Funding source distribution for all trials registered with clinicaltrials.gov. Left: trends in funding source distribution for all drug trials conducted after 1 Jan 1997 until 19 July 2017 (based on the start_date data field). Right: overall funding source distribution for all registered drug studies. NIH=National Institutes of Health

doi: 10.1136/bmj.k2130 | *BMJ* 2018;361:k2130 | the bmj

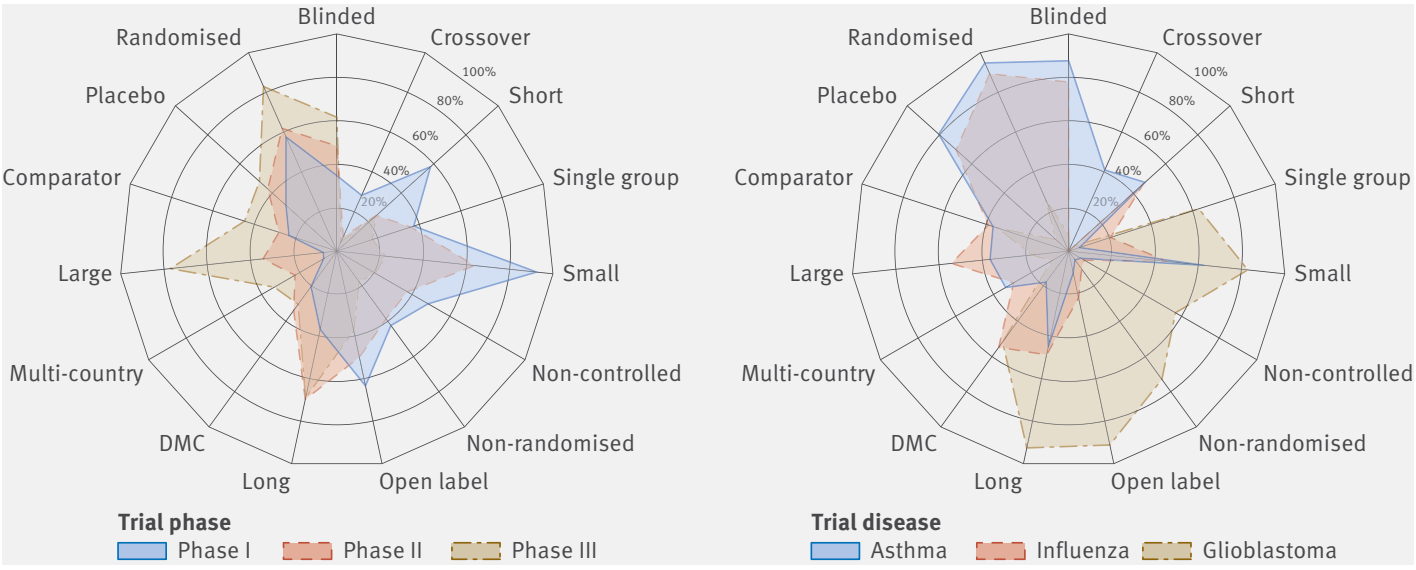

Fig 2 | Trial design properties. Each radar chart illustrates differences between three groups of trials (represented by coloured polygons) with respect to 15 trial protocol characteristics (individual radial axes). Each axis shows fraction of trials with given property, such as reported use of randomisation, blinding, or data monitoring committees (DMC). Left: trial characteristics by clinical phase. Right: characteristics of phase II treatment oriented trials for three representative diseases. Profile of glioblastoma (with many small non-randomised studies) is typical of oncology trials; see supplementary file 2. NIH=National Institutes of Health

by non-profit organisations in phase II has recently increased (fig 4). Notably, the use of blinding is still much lower in non-commercial than commercial trials. Additional analysis showed that the average numbers of trial eligibility criteria and outcome measures have increased over time, leading to a greater number of procedures performed in an average study, and hence to an increase in complexity of trials (supplementary file 2).

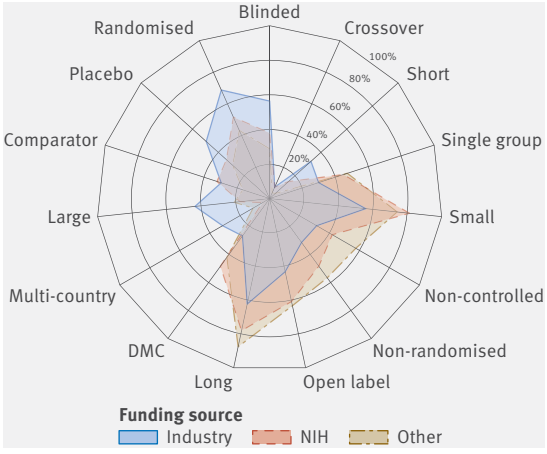

Fig 3 | Phase II trial properties by funding source. Radar chart illustrates differences between three groups of trials (represented by coloured polygons) with respect to 15 trial protocol characteristics (individual radial axes). Axis shows fraction of trials with given property, such as reported use of randomisation, blinding, or data monitoring committees (DMC). Detailed statistics across four funding categories and remaining clinical phases are available in supplementary file 4. NIH=National Institutes of Health

Dissemination of trial results and characteristics of published trials

Of the 45 620 completed trials, 27 835 were efficacy studies in phases II-IV covered by the mandate for results publication (supplementary file 1). Of these, only 15 084 (54.2%) have disclosed their results publicly. Consistent with previously reported trends,[38 39] dissemination rates were lower for earlier clinical phases (not covered by the publication mandate), as only 23.4% (2504/10 691) of trials in phases 0-I had publicly disclosed results (supplementary file 2).

Detailed analysis of phase II-IV trials showed that only 3822 (25.2%) of published studies disclosed their results within the required period of 12 months after completion. The median time to first public reporting, whether through direct submission to clinicaltrials.gov or publication in a journal, was 18.6 months. Dissemination lag was smaller for results submitted to clinicaltrials.gov (median 15.3 months) than for results published in a journal (median 23.9 months).

Of 27 835 phase II-IV studies, 10 554 (37.9%) disclosed the results through structured submission to clinicaltrials.gov, 8338 (29.95%) through a journal article, and 3808 (13.7%) through both resources. Scientific articles were published in 1454 titles of diverse nature and varying impact factor. These ranged from prestigious general medical journals (*BMJ*, *JAMA*, *Lancet*, *New England Journal of Medicine*) to narrower focus, specialist journals with no impact information available.

Journal publication status depended on study design. Trials reporting the use of randomisation were more likely to be published than non-randomised studies (35.0% (6895/19 711) *v* 18.2% (1408/7748)), and large trials enrolling more than 100 participants were more likely to be published than studies with less

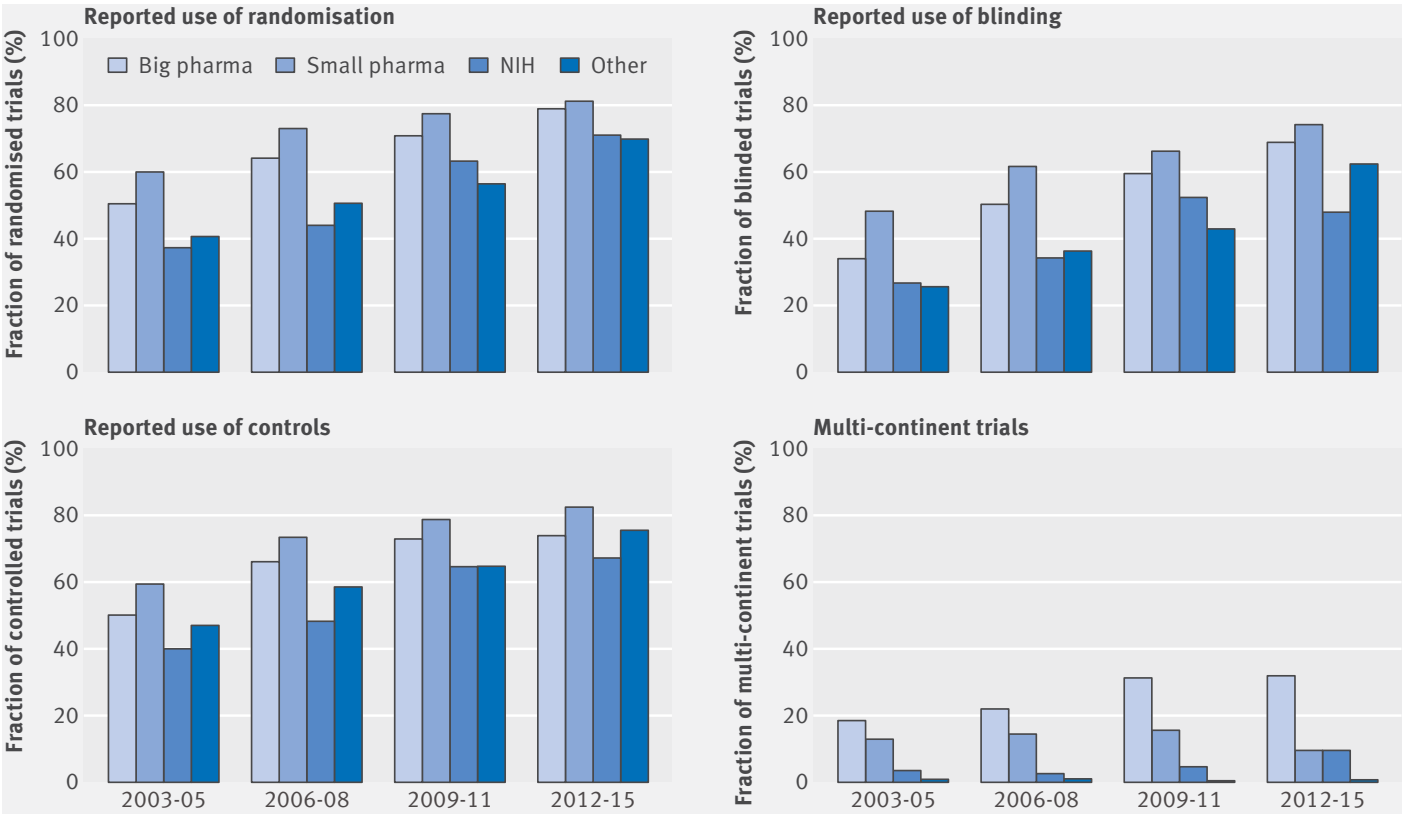

Fig 4 | Changes in clinical trial design over time. Plots compare characteristics of all phase II trials (regardless of completion status) divided into four temporal subsets based on their starting year. Additional results are available in supplementary file 2. NIH=National Institutes of Health

than 100 participants (39.3% (5541/14 106) *v* 20.5% (2753/13 422)).

In general, studies whose results were disclosed in the form of a scientific publication were more likely to follow the ideal randomised clinical trial design paradigms than were those whose results were submitted to clinicaltrials.gov only (fig 5). A similar trend was observed for trials with results published in higher versus lower impact journals. For instance, 89.6% (1009/1126) of trials published in journals with an H-index above 400 reported the use of randomisation, compared with 79.0% (1223/1548) of trials published in journals with an H index below 50 and 60.7% (5459/8999) of trials whose results were submitted to clinicaltrials.gov only (fig 5 and supplementary file 3). The characteristics of trials disclosed solely through submission to clinicaltrials.gov did not substantially differ from those not disseminated in any form (supplementary files 2 and 3).

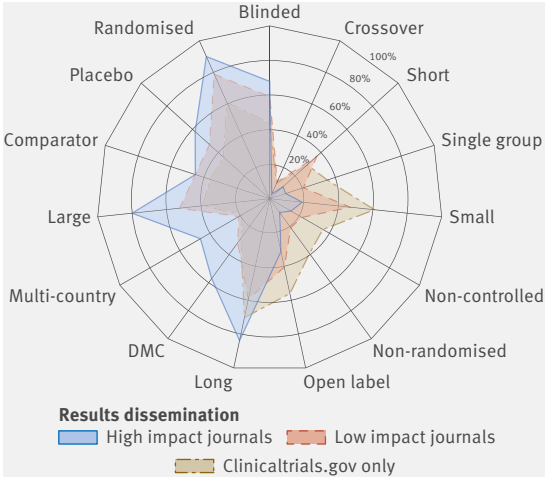

Fig 5 | Changes in clinical trial design over time. Plots compare characteristics of all phase II trials (regardless of completion status) divided into four temporal subsets based on their starting year. Additional results are available in supplementary file 2. NIH=National Institutes of Health

Trial dissemination rates across funder types

Trial dissemination rates varied among the different funding source categories (table 2 in supplementary file 2). In general, industry funded studies were more likely to disclose results than were non-commercial trials (59.3% (10 444/17 627) *v* 45.3% (4555/10 066)) of all completed phase II-IV studies). They were also more likely to report the results on clinicaltrials.gov (45.6% (8041) *v* 24.4% (2457)) and had a shorter time to reporting (median 14.5 *v* 18.0 months). Notably, only a small difference was observed for journal publication rates (31.0% (5460) *v* 28.1% (2827)). Additional analysis showed substantial differences across more granular funder categories (fig 6, top left). Among non-profit funders, a gap existed between NIH funded studies and those funded by other academic and non-profit organisations, with 60.0% (1451/2417)

doi: 10.1136/bmj.k2130 | *BMJ* 2018;361:k2130 | the bmj

and 40.6% (3104/7649) of trials being disclosed from the two categories, respectively. Similarly, among the industry funded clinical trials, big pharma had a higher result dissemination rate than small pharma (66.7% (7681/11 508) *v* 45.2% (2763/6119) of studies); see table 2 in supplementary file 2 for details.

In addition to overall reporting rates, the analysis also highlighted differences in the nature and timing of dissemination of results across trial funding sources (fig 6). Big pharma had the highest dissemination rate, both in terms of clinicaltrials.gov submission (53.2% (6122/11 508) of studies) and journal publications (34.8% (4005)). By contrast, funders classified as "other" showed the lowest percentage of trials published on clinicaltrials.gov (15.9% (1214/7649)). Organisations from this category disclosed trial results primarily through journal articles (29.1% (2229) of all studies) and were the only funder type with fewer trials reported through clinicaltrials.gov submission than through journal publication. The time to reporting trials at clinicaltrials.gov was fastest for industry funded trials and slowest for studies funded by organisations classed as "other;" publication lag for journal articles showed a reverse profile (see figure 6, bottom left and right, and table 2 in supplementary file 2 for detailed statistics).

Analysis of the fraction of trials with results disclosed within the required 12 months after completion showed that dissemination rates have been steadily growing with time, with the largest increase observed after 2007, when the FDA Amendments Act came into effect.[9] Further analysis showed that the observed growth was largely driven by increased timely reporting on clinicaltrials.gov, most evident for studies funded by big pharma and NIH (supplementary file 2).

Trial dissemination rates across disease categories

Disease areas with highest dissemination of results were autoimmune diseases and infectious diseases, whereas neurology and oncology showed the lowest dissemination rates (fig 7). As shown on the left side of figure 7, these differences were largely driven by differential trends in journal publication, with less variability observed for rates of result reporting

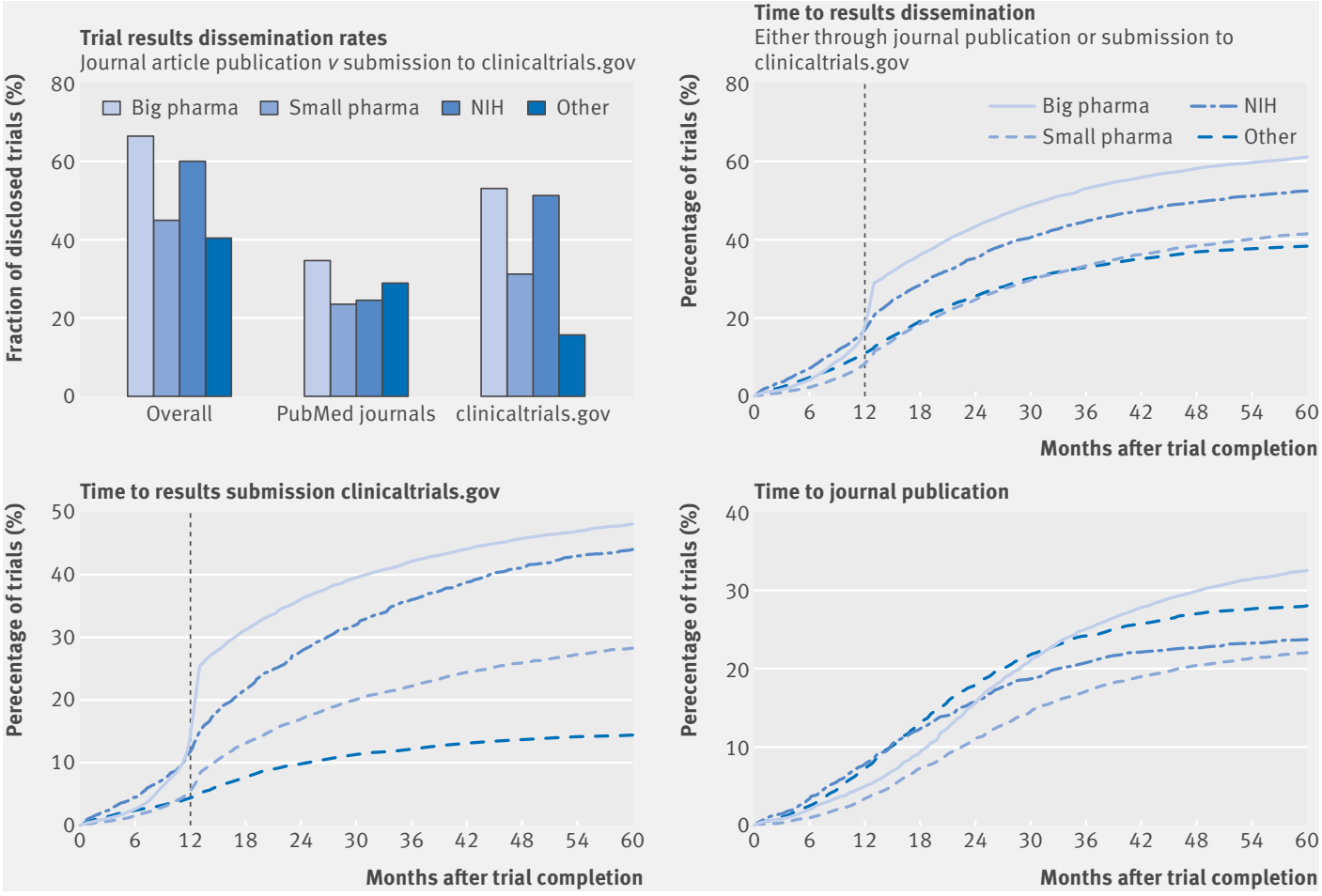

Fig 6 | Trial results dissemination rates by trial funding category. Top left: trial dissemination rates by funder type: overall dissemination rate, journal publications, and submission of structured results to clinicaltrials.gov. Top right: time to first results dissemination (whether through submission to clinicaltrials.gov results database or publication in journal article). Vertical line indicates 12 month deadline for reporting mandated by Food and Drug Administration Amendments Act. Bottom left: time to results reporting on clinicaltrials.gov. Bottom right: time to journal article publication. NIH=National Institutes of Health

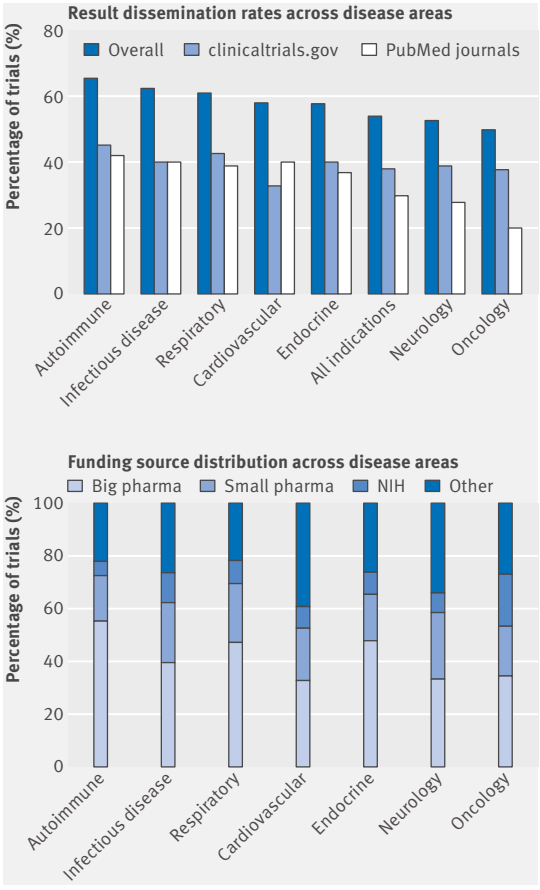

Fig 6 | Trial results dissemination rates and funding source distribution across seven medical specialties. Top: for each specialty, plot shows fraction of completed phase II-IV trials whose results were publicly disseminated, either as journal article or through direct submission to clinicaltrials.gov results database. Bottom: funding source distribution by medical specialty. NIH-National Institutes of Health

on clinicaltrials.gov. Oncology in particular had consistently the lowest journal publication rate across all clinical phases and funder categories.

The two disease categories with the highest fraction of published trials (autoimmune and infectious diseases) were also previously associated with highest likelihood of new drug approval (12.7% (n=549) and 16.7% (537), respectively).[31] Similarly, the two disease areas showing the lowest proportion of published trials also had the lowest drug approval rates: neurology (likelihood of 9.4%; n=986) and oncology (6.7%; n=1083).[31] Cardiovascular diseases had a relatively high publication rate but a low clinical trial success rate, with only 7.1% (n=426) of tested compounds eventually approved for marketing.[31]

Discussion

This analysis, enabled by automated data driven analytics, provided several important insights on the scope of research activity and the design, visibility, and reporting of clinical trials of drugs. Trial activity has grown and diversified, with a more diverse range of profit and non-profit organisations undertaking clinical research across a wide range of disease areas, and the analysis highlighted potentially important differences in the quality and dissemination of trials.

In terms of trial design, industry funded studies tended to use more robust methods than studies funded by non-profit institutions, although the methodological differences have been decreasing with time. Rates of dissemination of results varied substantially depending on the disease area as well as the type and size of funding institution. In general, the results of industry funded trials were more likely to be disclosed compared with non-commercial studies, although small drug companies lagged behind "big pharma." Similarly, trials with NIH funding more often disclosed their results than did those funded by other academic and non-profit organisations. Although overall dissemination rates have improved with time, the trend was largely driven by increased reporting of results on clinicaltrials.gov for big pharma and NIH funded studies. By contrast, the results of trials funded by other academic and non-profit organisations were primarily published as journal articles and were rarely submitted to clinicaltrials.gov. Finally, differences in dissemination rates across medical specialties were mainly due to differential journal publication rates ranging from 42% for autoimmune diseases to 20% for oncology.

Strengths and limitations of study

One of the major strengths of this work is the larger volume of clinical trials data and the greater depth and detail of the analysis compared with previous studies in this area. Many factors were analysed for the first time in the context of both quality of trial design and transparency of research. These included the distinction between big pharma and smaller industrial organisations, analysis of publication trends across distinct medical specialties, and differences between trials whose results were disseminated through submission to clinicaltrials.gov and publication in higher or lower impact journals. In addition, the study provided a novel method for visualising the characteristics of trials, which enables a rapid assimilation of similarities and differences across a variety of factors.

The study had several limitations related to the underlying dataset and the methods used to annotate and integrate the data. Firstly, it was limited to the clinicaltrials.gov database and hence did not include unregistered studies. However, given that registration of trials has been widely accepted and enforced by regulators and journal editors,[8 12] most of the influential studies will probably have been captured in the analysis.[40] Additional limitations were due to data quality factors such as missing or incorrect entries, ambiguous terminology, and lack of semantic standardisation.[14 16] Many trials were annotated as placebo or comparator controlled on the basis of information mined from their unstructured titles and descriptions, as the correct annotation in the

doi: 10.1136/bmj.k2130 | *BMJ* 2018;361:k2130 | the bmj

structured "arm type" field was often missing. Although we analysed a subset of annotations manually, our automatic workflow might still have misclassified some trials. In addition, owing to the non-standardised free text format of the data, we used only a simple count of eligibility criteria and outcome measures as an imperfect proxy for assessing the complexity of trials and the homogeneity of enrolled population.

Finally, our method for linking registry records to associated publications relied on the presence of a registry identifier (NCT number) in the text or metadata of a journal article. Its absence from a trial publication would lead to misclassified publication status and, consequently, underestimated publication rates. However, owing to the rigorousness of the search methods, the publication status is unlikely to have been misclassified in a systematic manner—for example, depending on study funding source or disease area. Hence, although absolute estimates might be affected, the comparative analysis based on automated record linkage is unlikely to have been biased towards one of the categories. Notably, reporting of registration numbers is encouraged by medical journals through ICMJE and the CONSORT checklist, and it can be argued that compliance is part of investigators' responsibility to ensure that their research is transparent and discoverable.[20 35]

Comparison with other studies

In addition to providing novel contributions, our analysis further confirmed and extended several findings reported previously by others. Notably, larger sample size, wide inclusion criteria, and additional mappings allowed us to calculate more accurate statistics and ask previously unexplored questions in an unbiased way. For instance, Bala et al showed that trials published in higher impact journals have larger enrolment size on the basis of a sample of 1140 reports of randomised clinical trials and manual searches.[41] Our study extended this analysis to the entire clinicaltrials.gov registry (not only randomised trials) and many additional trial properties beyond enrolment size, including reported use of data monitoring committees, duration of the studies, and number of countries. Other previously reported trends confirmed and further examined in this study included lower reporting rate observed for earlier clinical phases and academic sponsors,[38 39] longer dissemination lag for journal publications,[42] and overall improvements in dissemination of results.[43]

Several previous studies investigated the transparency of clinical research by using various subsets of clinicaltrials.gov and diverse approaches for classifying study results as published,[17 20 36-39 44 45] summarised in detail elsewhere.[46 47] General trends shown by our study were within estimates reported previously, and our estimate of 54% for the overall trial dissemination rate was in agreement with two recent systematic reviews.[46 47] Our study is methodologically most similar to the work by Powell-Smith and Goldacre,[20] with one important distinction: we did not restrict the analysis to trials sponsored by organisations with more than 30 studies, as this would exclude small drug companies and research centres and bias the results towards large institutions that routinely perform or fund clinical research. As a consequence, the analysis showed analogous trends but slightly lower overall reporting rates compared with that study.[20]

Meaning and implications of study

The higher methodological quality of industry funded trials may reflect the available financial and organisational resources, as well as better infrastructure and expertise in conducting clinical trials.[48 49] Differential objectives of the industrial and non-profit investigators (development of new drugs with a view to gaining a licence for market access versus mechanistic research and rapid publication of novel findings) may also have contributed to the observed differences.[48 50]

The analysis suggests that trials funded by large institutions that regularly conduct clinical research are more likely to have their results disseminated publicly. Although industry led reporting overall, large drug companies performed much better than smaller organisations, probably because many large companies have developed explicit disclosure policies in the recent years.[51-53] Similarly, NIH has developed specific regulations for disseminating the results of the studies it funds. The most recent such policy requires the publication of all NIH funded trials, including (unprecedentedly) phase I safety studies.[54] Although large trial sponsors seem to actively pursue better transparency,[38] smaller organisations, particularly those in the non-profit sector, were less responsive to the FDA Amendments Act publication mandate. Such institutions may be less able, or less motivated, to allocate the time and resources needed to prepare and submit the results, or they may be less familiar with the legislation.[38 40]

Our study suggests that the FDA mandate for dissemination of trial results and the establishment of the clinicaltrials.gov results database led to improved transparency of clinical research. The time to dissemination of results is shorter for clinicaltrials.gov than for traditional publication, and the database stores the results of many otherwise unpublished studies. Lack of journal publications might be explained by the reluctance of editors to publish smaller studies based on suboptimal design or studies with "negative" or non-significant outcomes that are unlikely to influence clinical practice.[41 55 56] Our analysis showed that journal publication varied depending on trial design and disease area, with certain medical specialties and smaller non-randomised studies more likely to be disseminated solely through clinicaltrials.gov. The difference was particularly evident for oncology. Only a fifth of completed cancer trials were published in a journal, perhaps owing to the high prevalence of small sample sizes, suboptimal trial designs, and negative outcomes in this field.[14 31] High rates of dissemination on clinicaltrials.gov, including for cancer trials and

small studies, suggests that this resource plays an important role in reducing publication bias. Thus, to assure better completeness of available evidence, the database should be used by clinical and drug discovery researchers in addition to literature reviews.

Importantly, automated analyses and methods for identification of yet to be published studies could be used by regulators and funders of clinical research to track trial investigators who failed to disclose the results of their studies. Such a technology would be particularly useful when coupled with automated notifications and additional incentives (or sanctions) for dissemination of results. Recently, the concept of automated ongoing audit of trial reporting was implemented through an online TrialTracker tool that monitors major clinical trial sponsors with more than 30 studies.[20] Improvements to the quality of clinical trial data could facilitate the development of robust monitoring systems in the future. Firstly, data quality should be taken into account when registering new trials and standardised nomenclature should be used whenever possible. Ideally, the user interface of the clinicaltrials.gov registry itself should be designed to encourage and facilitate normalisation. In addition, stricter adherence to the CONSORT requirement for including the trial registration number with each published article would improve the accuracy of automated mapping of registry records to associated publications.[40]

In addition to improved monitoring, new regulations might be needed to ensure better transparency, and it is important that small industrial organisations and academic centres are reached by these efforts as well. One possible solution was proposed by Anderson and colleagues[38]: journal editors from ICMJE could call for submission of results to clinicaltrials.gov as a requirement for article publication. Given the dramatic increase in trial registration rates that followed previous ICMJE policies,[57 58] such a call would be likely to improve the reporting of ongoing clinical research.[38]

Acknowledgements: We thank our colleagues from BenevolentAI, University College London, and Farr Institute of Health Informatics for much useful discussion and advice, particularly John Overington (currently Medicines Discovery Catapult), Nikki Robas, Bryn Williams-Jones, Chris Finan, and Peter Cox. We also thank the clinicaltrials.gov team for their assistance and advice.

Contributors: MZ, MD, ADH, and JH conceived and designed this study. MZ and MD acquired the data. MZ, MD, ADH, and JH analysed and interpreted the data. The initial manuscript was drafted by MZ; all authors critically revised the manuscript and approved its final version. MZ takes responsibility for the overall integrity of the data and the accuracy of the data analysis and attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. JH is the guarantor.

Funding: This study received no specific external funding from any institution in the public, commercial, or not for profit sectors and was conducted by Benevolent Bio as part of its normal business. ADH is supported by the UCL Hospitals NIHR Biomedical Research Centre and is an NIHR senior investigator.

Competing interests: All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not needed

Transparency: The manuscript's guarantor affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Data sharing: All the data used in the study are available from public resources. The dataset can be made available on request from the corresponding author.

1 Sibbald B, Roland M. Understanding controlled trials. Why are randomised controlled trials important? *BMJ* 1998;316:201. doi:10.1136/bmj.316.7126.201
2 Lindsay MA. Target discovery. *Nat Rev Drug Discov* 2003;2:831-8. doi:10.1038/nrd1202
3 Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ* 1996;312:71-2. doi:10.1136/bmj.312.7023.71
4 Woolf SH. The meaning of translational research and why it matters. *JAMA* 2008;299:211-3. doi:10.1001/jama.2007.26
5 Kendall JM. Designing a research project: randomised controlled trials and their principles. *Emerg Med J* 2003;20:164-8. doi:10.1136/emj.20.2.164
6 Simes RJ. Publication bias: the case for an international registry of clinical trials. *J Clin Oncol* 1986;4:1529-41. doi:10.1200/JCO.1986.4.10.1529
7 Yamey G. Scientists who do not publish trial results are "unethical". *BMJ* 1999;319:939A. doi:10.1136/bmj.319.7215.939a
8 US Food & Drug Administration. Food and Drug Administration Modernization Act (FDAMA) of 1997. https://www.fda.gov/RegulatoryInformation/LawsEnforcedbyFDA/SignificantAmendmentstotheFDCAct/FDAMA/default.htm.
9 US Food & Drug Administration. Food and Drug Administration Amendments Act (FDAAA). 2007. https://www.fda.gov/drugs/guidancecomplianceregulatoryinformation/guidances/ucm064998.htm.
10 Zarin DA, Tse T. Medicine. Moving toward transparency of clinical trials. *Science* 2008;319:1340-2. doi:10.1126/science.1153632
11 Zarin DA, Tse T, Williams RJ, Carr S. Trial Reporting in ClinicalTrials.gov—The Final Rule. *N Engl J Med* 2016;375:1998-2004. doi:10.1056/NEJMsr1611785
12 International Committee of Medical Journal Editors. Clinical trials: registration. http://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html.
13 World Health Organization. International Clinical Trials Registry Platform (ICTRP): primary registries. http://www.who.int/ictrp/network/primary/en/.
14 Hirsch BR, Califf RM, Cheng SK, et al. Characteristics of oncology clinical trials: insights from a systematic analysis of ClinicalTrials.gov. *JAMA Intern Med* 2013;173:972-9. doi:10.1001/jamainternmed.2013.627
15 Zarin DA, Ide NC, Tse T, Harlan WR, West JC, Lindberg DA. Issues in the registration of clinical trials. *JAMA* 2007;297:2112-20. doi:10.1001/jama.297.19.2112
16 Califf RM, Zarin DA, Kramer JM, Sherman RE, Aberle LH, Tasneem A. Characteristics of clinical trials registered in ClinicalTrials.gov, 2007-2010. *JAMA* 2012;307:1838-47. doi:10.1001/jama.2012.3424
17 Chen R, Desai NR, Ross JS, et al. Publication and reporting of clinical trial results: cross sectional analysis across academic medical centers. *BMJ* 2016;352:i637. doi:10.1136/bmj.i637
18 Hakala A, Kimmelman J, Carlisle B, Freeman G, Fergusson D. Accessibility of trial reports for drugs stalling in development: a systematic assessment of registered trials. *BMJ* 2015;350:h1116. doi:10.1136/bmj.h1116
19 Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of registered and published primary outcomes in randomized controlled trials. *JAMA* 2009;302:977-84. doi:10.1001/jama.2009.1242
20 Powell-Smith A, Goldacre B. The TrialsTracker: Automated ongoing monitoring of failure to share clinical trial results by all major companies and research institutions. *F1000Res* 2016;5:2629. doi:10.12688/f1000research.10010.1
21 Ross JS, Mulvey GK, Hines EM, Nissen SE, Krumholz HM. Trial publication after registration in ClinicalTrials.Gov: a cross-sectional analysis. *PLoS Med* 2009;6:e1000144. doi:10.1371/journal.pmed.1000144
22 Bradshaw J. Small drug companies grow as big pharma outsources. 2017. http://www.telegraph.co.uk/business/2017/01/22/top-50-privately-owned-pharma-companies-britain/.

23 Barden CJ, Weaver DF. The rise of micropharma. *Drug Discov Today* 2010;15:84-7. doi:10.1016/j.drudis.2009.10.001
24 Munos B. Lessons from 60 years of pharmaceutical innovation. *Nat Rev Drug Discov* 2009;8:959-68. doi:10.1038/nrd2961
25 Munos BH, Chin WW. How to revive breakthrough innovation in the pharmaceutical industry. *Sci Transl Med* 2011;3:89cm16. doi:10.1126/scitranslmed.3002273
26 Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical R&D. *Nat Rev Drug Discov* 2011;10:428-38. doi:10.1038/nrd3405
27 Thiers FA, Sinskey AJ, Berndt ER. Trends in the globalization of clinical trials. *Nat Rev Drug Discov* 2008;7:13. doi:10.1038/nrd2441.
28 Chondrogiannis E, Andronikou V, Tagaris A, Karanastasis E, Varvarigou T, Tsuji M. A novel semantic representation for eligibility criteria in clinical trials. *J Biomed Inform* 2017;69:10-23. doi:10.1016/j.jbi.2017.03.013
29 Coletti MH, Bleich HL. Medical subject headings used to search the biomedical literature. *J Am Med Inform Assoc* 2001;8:317-23. doi:10.1136/jamia.2001.0080317
30 Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet* 2006;7:119-29. doi:10.1038/nrg1768
31 Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. *Nat Biotechnol* 2014;32:40-51. doi:10.1038/nbt.2786
32 ClinicalTrials.gov. Glossary of common site terms. https://clinicaltrials.gov/ct2/about-studies/glossary.
33 ClinicalTrials.gov. ClinicalTrials.gov protocol registration data element definitions for interventional and observational studies. 2017 http://prsinfo.clinicaltrials.gov/definitions.html.
34 Ranking the Brands. Top 50 global pharma companies. 2017. https://www.rankingthebrands.com/The-Brand-Rankings.aspx?rankingID=370.
35 Huser V, Cimino JJ. Linking ClinicalTrials.gov and PubMed to track results of interventional human clinical trials. *PLoS One* 2013;8:e68409. doi:10.1371/journal.pone.0068409
36 Ramsey S, Scoggins J. Commentary: practicing on the tip of an information iceberg? Evidence of underpublication of registered clinical trials in oncology. *Oncologist* 2008;13:925-9. doi:10.1634/theoncologist.2008-0133
37 Lokker C, Haynes RB, Wilczynski NL, McKibbon KA, Walter SD. Retrieval of diagnostic and treatment studies for clinical use through PubMed and PubMed's Clinical Queries filters. *J Am Med Inform Assoc* 2011;18:652-9. doi:10.1136/amiajnl-2011-000233
38 Anderson ML, Chiswell K, Peterson ED, Tasneem A, Topping J, Califf RM. Compliance with results reporting at ClinicalTrials.gov. *N Engl J Med* 2015;372:1031-9. doi:10.1056/NEJMsa1409364
39 Prayle AP, Hurley MN, Smyth AR. Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: cross sectional study. *BMJ* 2012;344:d7373. doi:10.1136/bmj.d7373
40 Manzoli L, Flacco ME, D'Addario M, et al. Non-publication and delayed publication of randomized trials on vaccines: survey. *BMJ* 2014;348:g3058. doi:10.1136/bmj.g3058
41 Bala MM, Akl EA, Sun X, et al. Randomized trials published in higher vs. lower impact journals differ in design, conduct, and analysis. *J Clin Epidemiol* 2013;66:286-95. doi:10.1016/j.jclinepi.2012.10.005
42 Riveros C, Dechartres A, Perrodeau E, Haneef R, Boutron I, Ravaud P. Timing and completeness of trial results posted at ClinicalTrials.gov and published in journals. *PLoS Med* 2013;10:e1001566, discussion e1001566. doi:10.1371/journal.pmed.1001566
43 Miller JE, Wilenzick M, Ritcey N, Ross JS, Mello MM. Measuring clinical trial transparency: an empirical analysis of newly approved drugs and large pharmaceutical companies. *BMJ Open* 2017;7:e017917. doi:10.1136/bmjopen-2017-017917
44 Jones CW, Handler L, Crowell KE, Keil LG, Weaver MA, Platts-Mills TF. Non-publication of large randomized clinical trials: cross sectional analysis. *BMJ* 2013;347:f6104. doi:10.1136/bmj.f6104
45 Ross JS, Tse T, Zarin DA, Xu H, Zhou L, Krumholz HM. Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis. *BMJ* 2012;344:d7292. doi:10.1136/bmj.d7292
46 Chan AW, Song F, Vickers A, et al. Increasing value and reducing waste: addressing inaccessible research. *Lancet* 2014;383:257-66. doi:10.1016/S0140-6736(13)62296-5
47 Schmucker C, Schell LK, Portalupi S, et al, OPEN consortium. Extent of non-publication in cohorts of studies approved by research ethics committees or included in trial registries. *PLoS One* 2014;9:e114023. doi:10.1371/journal.pone.0114023
48 Laterre PF, François B. Strengths and limitations of industry vs. academic randomized controlled trials. *Clin Microbiol Infect* 2015;21:906-9. doi:10.1016/j.cmi.2015.07.004
49 Martin L, Hutchens M, Hawkins C, Radnov A. How much do clinical trials cost? *Nat Rev Drug Discov* 2017;16:381-2. doi:10.1038/nrd.2017.70
50 Ehlers MD. Lessons from a recovering academic. *Cell* 2016;165:1043-8. doi:10.1016/j.cell.2016.05.005
51 Novartis Public Affairs. Novartis position on clinical study transparency – clinical study registration, results reporting and data sharing. 2016. https://www.novartis.com/sites/www.novartis.com/files/clinical-trial-data-transparency.pdf.
52 Pfizer. Trial data & results. https://www.pfizer.com/science/clinical-trials/trial-data-and-results.
53 GlaxoSmithKline. Data transparency. 2014. https://www.gsk.com/en-gb/behind-the-science/innovation/data-transparency/.
54 National Institutes of Health. NIH policy on the dissemination of NIH-Funded clinical trial information. 2016. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-16-149.html.
55 Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev* 2009;1:MR000006.
56 Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ* 1997;315:640-5. doi:10.1136/bmj.315.7109.640
57 Laine C, Horton R, DeAngelis CD, et al. Clinical trial registration: looking back and moving ahead. *Lancet* 2007;369:1909-11. doi:10.1016/S0140-6736(07)60894-0
58 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. doi:10.1136/bmjopen-2015-008932

Web appendix: Supplementary material