



Check for updates

SOFTWARE TOOL ARTICLE

The TrialsTracker: Automated ongoing monitoring of failure to share clinical trial results by all major companies and research institutions [version 1; referees: 2 approved]

Anna Powell-Smith, Ben Goldacre

Evidence-Based Medicine Data Lab, Centre for Evidence-Based Medicine, Nuffield Department of Primary Health Care Sciences, University of Oxford, Oxford, UK

v1 First published: 03 Nov 2016, 5:2629 (doi: [10.12688/f1000research.10010.1](https://doi.org/10.12688/f1000research.10010.1))
Latest published: 03 Nov 2016, 5:2629 (doi: [10.12688/f1000research.10010.1](https://doi.org/10.12688/f1000research.10010.1))

Abstract

Background: Failure to publish trial results is a prevalent ethical breach with a negative impact on patient care. Audit is an important tool for quality improvement. We set out to produce an online resource that automatically identifies the sponsors with the best and worst record for failing to share trial results. **Methods:** A tool was produced that identifies all completed trials from clinicaltrials.gov, searches for results in the clinicaltrials.gov registry and on PubMed, and presents summary statistics for each sponsor online. **Results:** The TrialsTracker tool is now available. Results are consistent with previous publication bias cohort studies using manual searches. The prevalence of missing studies is presented for various classes of sponsor. All code and data is shared. **Discussion:** We have designed, built, and launched an easily accessible online service, the TrialsTracker, that identifies sponsors who have failed in their duty to make results of clinical trials available, and which can be maintained at low cost. Sponsors who wish to improve their performance metrics in this tool can do so by publishing the results of their trials.



This article is included in the [All trials matter](#) channel.

Open Peer Review

Referee Status:

Invited Referees		
	1	2
version 1		
published 03 Nov 2016	report	report
1 Andrew P. Prayle , University of Nottingham UK 2 James Hetherington , University College London UK, Jonathan Cooper , University College London UK Sinan Shi , University College London UK		

Discuss this article

[Comments](#) (3)

Corresponding author: Ben Goldacre (ben.goldacre@phc.ox.ac.uk)

How to cite this article: Powell-Smith A and Goldacre B. **The TrialsTracker: Automated ongoing monitoring of failure to share clinical trial results by all major companies and research institutions [version 1; referees: 2 approved]** *F1000Research* 2016, 5:2629 (doi: [10.12688/f1000research.10010.1](https://doi.org/10.12688/f1000research.10010.1))

Copyright: © 2016 Powell-Smith A and Goldacre B. This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: BG is funded by the Laura and John Arnold Foundation (LJAF) to conduct work on research integrity; APS is employed on this grant.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: BG has received research funding from LJAF, the Wellcome Trust, the NHS National Institute for Health Research, the Health Foundation, and the WHO. BG is co-founder of the AllTrials campaign on trials transparency. BG receives personal income from speaking and writing for lay audiences on the misuse of science. APS receives income as a freelance software developer.

First published: 03 Nov 2016, 5:2629 (doi: [10.12688/f1000research.10010.1](https://doi.org/10.12688/f1000research.10010.1))

Introduction

The results of clinical trials are used to make informed choices with patients about medical treatments. However, there is extensive and longstanding evidence that the results of clinical trials are routinely withheld from doctors, researchers, and patients. A current systematic review of all cohort studies following up registered trials, or trials with ethical approval, shows that approximately half fail to publish their results¹. Evidence from an earlier review shows that studies with “negative” or non-significant results are twice as likely to be left unpublished². Legislation, such as FDA Amendment Act 2007 (<http://www.fda.gov/Regulatory-Information/Legislation/SignificantAmendmentstotheFDCAct/FoodandDrugAdministrationAmendmentsActof2007/default.htm>), which requires trials to post summary results on clinicaltrials.gov within 12 months of completion, have been widely ignored, with a compliance rate of one in five^{3,4}. The FDA is entitled to impose fines of \$10,000 a day on those breaching this law, but has never yet done so^{5,6}. This public health problem has also been the subject of extensive campaigning. For example, the AllTrials campaign is currently supported by 89,000 individuals and 700 organisations, including major funders, professional bodies, patient organisations and government bodies (<http://www.alltrials.net/>).

Previous work suggests that some sponsors, companies, funders, and research sites may perform better than others^{5,7}. In any sector, audit of the best and worst performers can be used to improve performance, allowing those with a poor performance to learn from those doing better. To be effective, however, audit should be repeated, and ideally ongoing⁸.

All work on publication bias to date relies on a single sweep of labour-intensive manual searches^{9,10}, or a single attempt to automatically match registry entries to published papers using registry identification number¹¹. Manual matching comes at high cost and does not give ongoing feedback. We therefore set out to: develop an online tool that automatically identifies trials with unreported results; present and rank the prevalence of publication failure, broken down by sponsor; and maintain the service, updating the data automatically, so that companies and research institutes are motivated to improve their performance.

Methods

The methods used by the online tool are as follows. Raw structured data on all studies in clinicaltrials.gov are downloaded in XML format. Studies are kept if they: have a study type “interventional” (excluding observational studies); have a “status” of “completed”; have a completion date more than 24 months ago, and after Jan 1 2006; are phase 2, 3, 4, or “n/a” (generally a device or behavioural intervention); no application to delay results posting has been filed (ascertained from the *firstreceived_results_disposition_date* tag); are conducted by a sponsor who has sponsored more than 30 trials (to exclude trials conducted by minor sponsors and make the ranking in the tool more informative).

Results are then sought for all included studies, using two methods. First the tool checks for structured results posted directly in clinicaltrials.gov, ascertained by the presence of the *firstreceived_results_date* tag. Secondly, the tool searches for the *nct_id* (registry

ID number) of the trial in PubMed’s *Secondary Source ID* field. Since 2005, all trials with a registry ID in the body of the journal article text should have that ID replicated in this field (https://www.nlm.nih.gov/bsd/policy/clin_trials.html). However, since in our experience approximately 1.5% of PubMed records include a valid *nct_id* list in the abstract, but not the *Secondary Source ID* field, our tool additionally searches for this ID in the title or abstract text. We exclude results published before the completion date of the trial, or results that have the words “study protocol” in the title.

A final filter is then applied, with the aim of excluding publications reporting protocols or additional analysis and commentary, rather than trial results; after experimenting with the standard validated PubMed “therapy” filters (both broad and narrow) and a rudimentary search for “study protocol”, the former was used. A comparison of the three methods is reported in the accompanying iPython notebook [<https://github.com/ebmdatalab/trialstracker>]¹².

Accepting that an automated tool cannot produce results with the accuracy of a manual search, we also performed some rudimentary checks of the output of the automated search against existing manual search cohorts. The overall prevalence of unreported studies found by the tool was compared against three previous studies on publication bias. In addition, disparities on individual studies found to be unreported by the tool were compared against the underlying data from a recent publication bias cohort study conducted using clinicaltrials.gov data.

The output data is then shared through an interactive website at <https://trialstracker.ebmdatalab.net> allowing users to rank sponsors by number of trials missing, number of trials conducted, and proportion of trials missing. Users can click on a sponsor name to examine the number and proportion of trials completed and reported from each year for that sponsor. The site URL changes as users focus on each organisation’s performance, so that users can easily share insights into the performance of an individual company or institution. By default sponsors are sorted by the highest number of unreported trials, rather than the highest proportion, in order to initially focus on larger and more well-known organisations. The site is designed responsively to be usable on mobile, tablet or desktop devices.

For transparency and replication, all code for the tool, with comments and all data sources, is available as an iPython notebook¹². All software is shared as open source, under the MIT license. A full CSV is shared containing all data, including all studies before our filters are applied, allowing others to conduct additional analyses or sensitivity analyses with different filtering methods.

Results

The TrialsTracker tool was successfully built and is now running online at <https://trialstracker.ebmdatalab.net>. Sample screenshots are presented in [Figure 1](#) and [Figure 2](#).

Since Jan 2006, trial sponsors included in our dataset have completed 25,927 eligible trials, of which 11,714 (45.2%) have failed to make results available. [Table 1](#) to [Table 4](#) report the sponsors with the top five highest number of unreported trials, the

Who's not sharing their trial results?

Trials registered on [ClinicalTrials.gov](https://clinicaltrials.gov) should share results on the site shortly after completing, or publish in a journal. But many organisations [fail to report the results of clinical trials](#). We think [this should change](#). Explore our data (last updated October 2016) to see the universities, government bodies and pharmaceutical companies that aren't sharing their clinical trial results.

Trial sponsors

We've ranked the major trial sponsors with the most unreported trials registered on [ClinicalTrials.gov](https://clinicaltrials.gov). Click on a sponsor's name to find out whether it's getting better at reporting completed trials - or worse.

	Name of sponsor	Trials missing results	Total eligible trials	Percent missing
1	Sanofi	285	435	65.5%
2	Novartis Pharmaceuticals	201	534	37.6%
3	National Cancer Institute (NCI)	194	558	34.8%
4	Assistance Publique - Hôpitaux de Paris	186	292	63.7%
5	GlaxoSmithKline	183	809	22.6%
6	Mayo Clinic	157	312	50.3%
7	Yonsei University	139	194	71.6%
8	Seoul National University Hospital	131	207	63.3%

Trials by year

Since Jan 2006, **all major trial sponsors** completed 25,927 eligible trials and **haven't published results for 11,714 trials**. That means 45.2% of their trials are missing results.

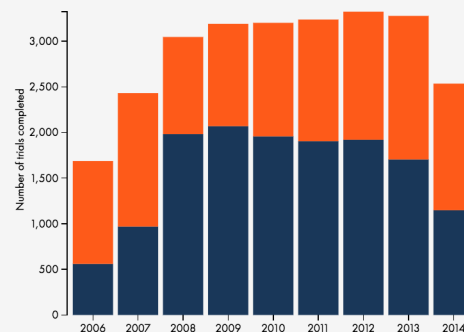


Figure 1. Screenshot, all trials. <https://trialstracker.ebmdatalab.net/>.

Who's not sharing their trial results?

Trials registered on [ClinicalTrials.gov](https://clinicaltrials.gov) should share results on the site shortly after completing, or publish in a journal. But many organisations [fail to report the results of clinical trials](#). We think [this should change](#). Explore our data (last updated October 2016) to see the universities, government bodies and pharmaceutical companies that aren't sharing their clinical trial results.

Trial sponsors

We've ranked the major trial sponsors with the most unreported trials registered on [ClinicalTrials.gov](https://clinicaltrials.gov). Click on a sponsor's name to find out whether it's getting better at reporting completed trials - or worse.

	Name of sponsor	Trials missing results	Total eligible trials	Percent missing
6	Mayo Clinic	157	312	50.3%
7	Yonsei University	139	194	71.6%
8	Seoul National University Hospital	131	207	63.3%
9	Alliance for Clinical Trials in Oncology	129	160	80.6%
10	Novartis	127	349	36.4%
11	University of British Columbia	126	174	72.4%
12	Merck Sharp & Dohme Corp.	125	612	20.4%
13	University of California, San Francisco	123	260	47.3%
14	University of Sao Paulo	122	193	63.2%

Trials by year

Since Jan 2006, **Mayo Clinic** completed 312 eligible trials and **hasn't published results for 157 trials**. That means 50.3% of its trials are missing results. See [all its completed trials on ClinicalTrials.gov](#).

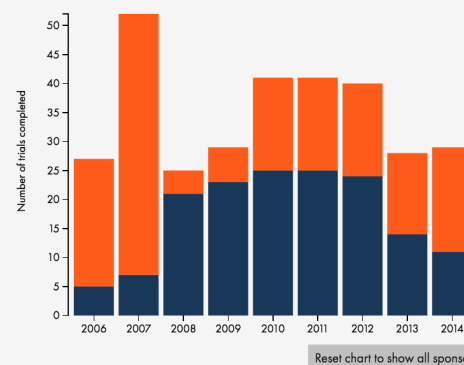


Figure 2. Screenshot, all trials by Mayo Clinic. <https://trialstracker.ebmdatalab.net/#mayo-clinic>.

Table 1. Top five sponsors with the highest number of missing results. TrialsTracker, 20/10/2016.

Name of trial sponsor	Trials missing results	Total eligible trials	Percent missing
Sanofi	285	435	66%
Novartis Pharmaceuticals	201	534	38%
National Cancer Institute (NCI)	194	558	35%
Assistance Publique - Hôpitaux de Paris	186	292	64%
GlaxoSmithKline	183	809	23%

Table 2. Top five sponsors with the highest number of eligible trials. TrialsTracker, 20/10/2016.

Name of trial sponsor	Trials missing results	Total eligible trials	Percent missing
GlaxoSmithKline	183	809	23%
Merck Sharp & Dohme Corp.	125	612	20%
National Cancer Institute (NCI)	194	558	35%
Novartis Pharmaceuticals	201	534	38%
Pfizer	62	471	13%

Table 3. Top five sponsors with the greatest proportion of missing trials. TrialsTracker, 20/10/2016.

Name of trial sponsor	Trials missing results	Total eligible trials	Percent missing
Ranbaxy Laboratories Limited	35	35	100%
Nanjing Medical University	32	35	91%
Rambam Health Care Campus	27	30	90%
Isfahan University of Medical Sciences	44	49	90%
City of Hope Medical Center	39	44	89%

Table 4. Top five sponsors with lowest proportion of missing trials. TrialsTracker, 20/10/2016.

Name of trial sponsor	Trials missing results	Total eligible trials	Percent missing
Shire	0	96	0%
Colgate Palmolive	1	32	3%
Bristol-Myers Squibb	5	115	4%
Eli Lilly and Company	15	292	5%
Johnson & Johnson Pharmaceutical Research & Development, L.L.C.	3	58	5%

highest number of eligible trials, the highest proportion of unreported trials, and the lowest proportion of unreported trials. In total, 2390/8799 (27.2%) trials with sponsors classed as “industry” were identified as unreported; 122/470 (26.0%) trials with sponsors classed as “US Fed” were identified as unreported; 361/996 (36.2%) trials with sponsors classed as “NIH” were identified as unreported; 8841/15662 (56.4%) trials with sponsors classed as “other” were identified as unreported. We find that 8.7 million patients were enrolled in trials that are identified as unreported.

Checks for consistency with previous work

A previous paper automatically matching registry entries to PubMed records and clinicaltrials.gov results found 55% had no evidence of results¹¹, consistent with our overall findings. A previous manual audit (of which BG is co-author) found 56% of trials conducted in the University of Oxford reported results; our method also found 56% for the same institution⁹. A previous manual audit examined 4347 trials across 51 academic medical centres⁷. We compared their individual study data against ours and found that 2562 trials (62.6%) in their cohort were also in ours, but note that their study only represented 2% of our total cohort. For studies in both cohorts we found 60% reported results, while they found 66%. Of studies in both cohorts: 1149 were found “reported” by both; 534 studies were found “unreported” by both; 497 were found “reported” by their method and “unreported” by ours; 382 were found “unreported” by theirs and “reported” by ours.

Discussion

The tool was successfully built, and is now fully functional online. We found non-publication rates consistent with those from previous work using manual searches, and reasonable consistency with individual study matches from a previous manual cohort. A wide range of publication failure rates were apparent in the data.

Strengths and weaknesses

Our tool is the first to provide live ongoing interactive monitoring of failure to publish the results of clinical trials. The method of automatic matching has strengths and weaknesses. It can be run automatically, at a lower unit cost than a manual search, and therefore allows coverage of more trials than any traditional cohort study. It also permits repeated re-analysis at minimal additional marginal cost compared to a manual search.

In corollary, the efficiency of automatic matching also brings challenges around specificity and sensitivity. Firstly, there may be false adjudications of non-publication, i.e. if a trial’s results paper does not include its registry identifier. However, since 2005 all major medical journals (through the International Committee of Medical Journal Editors; <http://icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html>) have required trials to be registered, and all trials should include their registry ID in the text. Therefore, in our view, the responsibility for results being undiscoverable, when the registry ID is not included by the trialists, lies solely with the trialists; research that is hard to discover is not transparently reported. We hope that in the future better methods for probabilistic record linkage will also be available for wider use¹³. Secondly, there may be false positives, where a study identified through ID matching and then filtered, is in fact not reporting results. We have used standard filters to account for

this, and we are keen to improve our method in the light of concrete constructive feedback. Our checks for consistency against overall prevalence findings and individual study data from previous research to a large extent exclude gross errors in prevalence figures.

Notably there are specific additional methods for linking clinicaltrials.gov records to PubMed records that we tried and rejected. Some trials have a link to a PubMed record directly in the clinicaltrials.gov results_reference tag, which ClinicalTrials documentation (<https://prsinfo.clinicaltrials.gov/definitions.html>) suggests indicates results from a publication. We found 2263 eligible trials had such tags, but no summary results on ClinicalTrials.gov. However, on manual examination, we found these are often erroneous, and commonly report results of unrelated studies from several years previously. In discussion, clinicaltrials.gov staff confirmed that this field is neither policed nor subject to substantial editorial control (personal communication with Annice Bergeris).

Context of other findings

Our findings are consistent with previous work on publication bias¹, finding that approximately half of trials fail to report results. Previous studies have used 2007 as their start date for expecting results to be made available, reflecting the FDA Amendment Act 2007. We did not use this date, as this legislation has been widely ignored^{5,6}, and because we regard sharing results as an ethical obligation, not a legal one. Our methods accept results posting at any time after study completion, and any sponsor posting results for any trial since 2006 will find their results improve in our live data.

Policy implications

We have previously argued that live ongoing monitoring of trials transparency will help to drive up standards, especially if this information is used by clinicians, policymakers, ethics committees, regulators, patients, patient groups, healthcare payers, and research funders, to impose negative consequences on those who engage in the unethical practice of withholding trial results from doctors, researchers, and patients¹⁴. Recent comments by US Vice President Joe Biden threatened to withhold financial support from publicly-funded researchers who fail to report clinical trial results, suggesting some consequences may arise⁶. We would be happy to collaborate or work with organisations seeking to get a better understanding of their own failure to publish, and wishing to act on this data.

We have also previously argued that medicine has an “information architecture” problem; all publicly accessible documents and data on all clinical trials should be aggregated and indexed for comparison and gap identification, and that good knowledge management and better use of trial identifiers will facilitate this¹⁵. At present, medicine faces serious shortcomings in this area. With 75 trials and 11 systematic reviews being published every day on average¹⁶ better knowledge management must be a priority.

Future research

We have shared all our underlying data so that others can explore in detail non-publication for specific studies, interventions, companies, funders, sponsors, or institutions that interest them. We believe that research work on research methods and reporting should go beyond identifying the overall prevalence of problems, and identify individual people and organisations who are performing poorly, in order to both support and incentivise them to improve. That is only possible with ongoing monitoring and feedback on individual studies, an approach we have taken on other projects such as COMPare^{17,18}. We hope that others will also pursue this model of audit and feedback, and assess its impact on performance.

Conclusions

We have designed, built, and launched an easily accessible online service that identifies sponsors who have failed in their duty to make results of clinical trials available.

Software availability

Website available at: <https://trialstracker.ebmdatalab.net>

Latest source code: <https://github.com/ebmdatalab/trialstracker>

Archived source code as at the time of publication: DOI: [10.5281/zenodo.163522](https://doi.org/10.5281/zenodo.163522)¹²

License: MIT license

Author contributions

BG conceived the project; both authors developed the analyses, trial matching and filtering methods; APS wrote the data-analysis script and built the interactive website; BG drafted the manuscript; both authors revised the manuscript; both authors are guarantor.

Competing interests

BG has received research funding from LJAF, the Wellcome Trust, the NHS National Institute for Health Research, the Health Foundation, and the WHO. BG is co-founder of the AllTrials campaign on trials transparency. BG receives personal income from speaking and writing for lay audiences on the misuse of science. APS receives income as a freelance software developer.

Grant information

BG is funded by the Laura and John Arnold Foundation (LJAF) to conduct work on research integrity; APS is employed on this grant.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements:

We are grateful for constructive discussions on design and impact with Jess Fleminger, Carl Heneghan and Sile Lane.

References

1. Schmucker C, Schell LK, Portalupi S, *et al.*: **Extent of Non-Publication in Cohorts of Studies Approved by Research Ethics Committees or Included in Trial Registries.** *PLoS One.* 2014; 9(12): e114023.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Song F, Parekh S, Hooper L, *et al.*: **Dissemination and publication of research findings: an updated review of related biases.** *Health Technol Assess.* 2010; 14(8): iii,ix–xi, 1–193.
[PubMed Abstract](#) | [Publisher Full Text](#)
3. Anderson ML, Chiswell K, Peterson ED, *et al.*: **Compliance with results reporting at ClinicalTrials.gov.** *N Engl J Med.* 2015; 372(24): 1031–1039.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Prayle AP, Hurley MN, Smyth AR: **Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: cross sectional study.** *BMJ.* 2012; 344: d7373.
[PubMed Abstract](#) | [Publisher Full Text](#)
5. **Patients endangered as law is ignored.** *STAT.* 2015.
6. **Joe Biden: Agencies don't report clinical trials should lose funds.** *STAT.* 2016.
[Reference Source](#)
7. 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.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Benjamin A: **Audit: how to do it in practice.** *BMJ.* 2008; 336(7655): 1241–1245.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Thompson AC, Petit-Zeman S, Goldacre B, *et al.*: **Getting our house in order: an audit of the registration and publication of clinical trials supported by the National Institute for Health Research Oxford Biomedical Research Centre and the Musculoskeletal Biomedical Research Unit.** *BMJ Open.* 2016; 6(6): e009285.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Miller JE, Korn D, Ross JS: **Clinical trial registration, reporting, publication and FDAAA compliance: a cross-sectional analysis and ranking of new drugs approved by the FDA in 2012.** *BMJ Open.* 2015; 5(5): e009758.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Huser V, Cimino JJ: **Linking ClinicalTrials.gov and PubMed to track results of interventional human clinical trials.** *PLoS One.* 2013; 8(7): e68409.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Powell-Smith A: **ebmdatalab/trialstracker: First release [Data set].** *Zenodo.* 2016.
[Data Source](#)
13. Bashir R, Dunn AG: **Systematic review protocol assessing the processes for linking clinical trial registries and their published results.** *BMJ Open.* 2016; 6(10): e013048.
[PubMed Abstract](#) | [Publisher Full Text](#)
14. Goldacre B: **How to Get All Trials Reported: Audit, Better Data, and Individual Accountability.** *PLoS Med.* 2015; 12(4): e1001821.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Goldacre B, Gray J: **OpenTrials: towards a collaborative open database of all available information on all clinical trials.** *Trials.* 2016; 17: 164.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Bastian H, Glasziou P, Chalmers I: **Seventy-five trials and eleven systematic reviews a day: how will we ever keep up?** *PLoS Med.* 2010; 7(9): e1000326.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. Goldacre B: **Make journals report clinical trials properly.** *Nature.* 2016; 530(7588): 7.
[PubMed Abstract](#) | [Publisher Full Text](#)
18. Goldacre B, Drysdale H, Slade E, *et al.*: **The COMPare Trials Project.** *COMPare.*
[Reference Source](#)

Open Peer Review

Current Referee Status:



Version 1

Referee Report 19 December 2016

doi:10.5256/f1000research.10786.r17404



James Hetherington¹, Sinan Shi², Jonathan Cooper²

¹ Research Software Development Group, University College London, London, UK

² University College London, London, UK

The authors published an online ranking system which illustrates how the major sponsors share their clinical trial information, in particular through reporting on completed trials. This research offers a new way to automatically identify and match trials registered on ClinicalTrials.gov with their published results in both the ClinicalTrials.gov trial registry and abstracts or metadata of publications (indexed in Pubmed). This automated process can result in a much more frequent update and provide more precise information to the public, in part by encouraging more accessible reporting.

In this review, we would like to focus our comments on the author's data processing and software. The authors have provided a code repository containing their website along with some Python code related to the data analysis process. The latter comprises a clear and straightforward IPython notebook detailing all the data analysis steps, including raw data processing, missing trial identification, and validation against other studies. We found it is an intuitive way to present works of this scale, although as discussed later we would like to suggest more modularization. In general, the code is understandable and easy to read. Both unit tests and behavioural tests are included to give more confidence in its reliability. We were able to re-run the entire IPython notebook with only some minor modifications.

We do have some minor comments and suggestions regarding the coding quality and reproducibility aspects of this project.

- We have noticed that the XML parsing and Pubmed data extraction parts break easily due to variations in the source files or network problems. It would therefore be beneficial to make these two parts into functions with associated unit tests to ensure the correctness and robustness of the code.
- Compounding the problem, these parts also take a very long time to compute. We left the program running for several days trying to update the trial-abstract database, only to have it fail part-way through. Further incremental updating mechanisms would help greatly here, for instance adding an extra column to the database to register the last search date so that recently searched entries will not be queried again.
- One hopes that the 'live' website is updated from time to time with more recent results. It would be nice to have details on how frequently this happens - is it an automated process?

- The current data on Github have some small differences compared to the results presented in the paper. We can fully understand that the data in the repository should be updated, and the development is an ongoing process. However, it would have been good from an audibility point of view to make the data which have been used for the paper available. For instance, the specific git commit id used for the paper could be given in the paper itself and the repository's README.
- A requirements.txt is provided in the source code to facilitate installing the project's dependencies, however, not all of the dependencies are on the list. Changes in recent versions of some of these cause the code to break. Please specify all the dependencies (even indirect ones) including the versions used in the requirements.txt file. We have submitted a pull request with the list we found worked.

Overall, the new tool offered by the authors enables more frequent and larger-scale identification of whether trials have been reported. Their code is clear and reflects the methodology faithfully. This tool will help in the push for improving clinical trial transparency.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 21 November 2016

doi:10.5256/f1000research.10786.r17811



Andrew P. Prayle

Division of Child Health, University of Nottingham, Nottingham, UK

Title and abstract

The title is appropriate and discusses the content of the paper in one sentence. The abstract starts generally, drills down into the methods concisely and discusses the contribution to the literature which this manuscript and software project appropriately.

Article content

Powell-Smyth and Goldacre report on a piece of work which will make a substantial contribution to the clinical trials enterprise.

They have developed an open source web application which automatically takes data from the US based ClinicalTrials.gov registry and searches for results (either summary results on ClinicalTrials.gov or an abstract on PubMed). The software then ranks study sponsors by the proportion of trials which have reported results.

This approach is novel in its approach to on-line availability of data. This means that the dataset is easily searchable through a web based application. Automated systems have been explored in the past (e.g. Huser *et al* 2013), as have manual searches (Tompson 2016), and the results of the automated system presented appear consistent with these.

I have reviewed the online web based software, this is simple to use and demonstrates the ability of the approach to hold institutions which sponsor research to account, by summarising their contribution of

results to the clinical trial literature.

The central contribution is an automated system for determining if a trial registered on ClinicalTrials.gov has published summary records on clinicaltrials.gov, or has an abstract indexed on PubMed. The work hinges on whether their automated system can in fact do this. The authors make a persuasive case that they are able to find summary results and abstracts where these have been published. They provide what they have said they can do in the on-line Jupyter notebook. Additionally, the open source code in the Github repository is straightforward to read, and supports their case. Finally, I downloaded the full dataset and explored it, and in the cases which I looked at their spreadsheet had correctly identified completed trials and the accompanying Pubmed abstract.

Therefore, although there may be a few trials which have been misclassified, I think that the methods used appear very robust. Additionally, if trials have been misclassified, the authors give suggestions of how to adjust this through changes to the journal entries on Pubmed, or through summary results on Pubmed.

In the discussion the strengths and limitations of their automated approach are carefully elaborated upon. The key strength is that a large proportion of the clinical trials landscape is included in their study. The limitation is of course that automated analysis may incorrectly label some trials as unreported when in fact they are, but my assessment of their raw data is that this must be infrequent as I have not been able (in and admittedly unscientific sample obtained by scrolling through the raw data, and looking at trials which I am familiar with if I see them) to identify such a case.

Conclusions

The authors state that they present this work to aim to improve the clinical trials landscape in terms of the 'information architecture' of missing results. I believe that we should take this work at face value as a genuine, innovative approach which is trying to improve the problem of non-reporting, by giving transparency of reporting at the study sponsor level. It is reported carefully. The data presented back up the case for a clear need for improved trial reporting.

Data

This study is an exemplar of how to publish reproducible research. The data and code and extensive documentation are available and free to download and explore. My only suggestion is to have a second repository in case GitHub disappears.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: I was joint co-author on a paper which was cited by this manuscript. Dr Goldacre has cited my work in a statement given to a House of Commons select committee, and has given a statement to me in support of an application which I made to the University of Nottingham which supported the impact of my work in this field.

Discuss this Article

Version 1

Reader Comment 23 Nov 2016

Adam Jacobs, Associate Director, Biostatistics, Premier Research, UK

Since my previous comment, I have looked into the under-estimation of disclosure some more, using a more representative sample.

My estimate is that a little over half of the trials identified as "undisclosed" by the Trials Tracker are in fact disclosed, so that the real proportion of undisclosed trials is not 45%, but 21% (12% for industry trials and 26% for non-industry trials).

For details of how I estimated those figures, please see my [blogpost](#) on the subject.

Competing Interests: As stated previously

Reader Comment 14 Nov 2016

Tamas Ferenci, Obuda University, Hungary

A quick update: I also checked AstraZeneca (this time a completely randomly chosen company), and the story seems to be similar. They have an own disclosure site (<http://www.astrazenecaclinicaltrials.com/ST/Submission/Search>), and from the 68 AstraZeneca trials that TrialsTracker reports as overdue, 38 (55.9%) can be found on that site. (But, they weren't even uploaded to clinicaltrials.gov, much less published in a PubMed-indexed journal.)

Let me again repeat that this represents a very bad practice, as these results will have practically zero visibility, they likely won't be found by researchers... but it is also not fair to call them "unreported" either.

So while TrialsTracker's results are invalid for AstraZeneca too, they also draw attention to the fact, how bad is the indexing, dissemination of these results. (For reasons that are entirely unclear to me, again, at least as far as the uploading to clinicaltrials.gov is concerned.)

Limitations of this remark are the same as my earlier one.

One might wonder whether it'd make sense to "correct" the results of TrialsTracker (i.e. Sanofi or AstraZeneca) based on these findings. But it likely doesn't make sense – even apart from the fact that what it now measures is meaningful, even if it is not "non-publication" – because that would make different sponsor's results incomparable (whether they're manually corrected or not). In this situation it is better to be uniformly wrong than wrong sometimes and correct other times, making comparison impossible.

Transparency: The R code is now more complicated, as we cannot list all trials, we have to do a search. Also the presence of results is not given unambiguously; I assumed that a trial has results if an attachment is uploaded which has the string "CSR" in its name.

```
library( rvest )
```

```
all <- read.csv( "all.csv", stringsAsFactors = FALSE )
```

```
AZ.url <- "http://www.astrazenecaclinicaltrials.com/ST/Submission/Search"
```

```
AZOverdue <- subset( all, lead_sponsor=="AstraZeneca"&is_overdue=="True" )$nct_id
```

```
AZOverdueReport <- sapply( AZOverdue, function( x ) {
```

```
  s <- html_session( AZ.url )
```

```

AZ.url <- submit_form( s, set_values( html_form( html_session( AZ.url ) )[[ 1 ]], searchString = x ) )$url
if ( AZ.url=="http://www.astrazenecaclinicaltrials.com/ST/Submission/SearchResults" )
  return( FALSE )
else {
  AZReported <- html_text( html_nodes( read_html( AZ.url ), '#attachmentsTab' ) )
  if( AZReported=="No attachments posted." )
    return( FALSE )
  else
    return( grepl( "CSR", AZReported ) )
}
} )
table( AZOverdueReport )

```

Competing Interests: None.

Reader Comment 12 Nov 2016

Tamas Ferenci, Obuda University, Hungary

First, I'd like to congratulate Anna Powell-Smith, Ben Goldacre and the entire team at EBM Data Lab for this highly relevant, interesting and – computationally – exciting project. The results have utmost importance in my opinion, but it also means that great care should be taken to check their validity.

I (completely independently from Adam Jacobs) also found results pertaining to Sanofi quite strange. Not because I have any special information specifically about Sanofi's trials, but it was a so obvious outlier. So I started some manual experimentation, and I also quickly found the same site Adam Jacobs did (http://en.sanofi.com/Innovation/clinical_trials/our_commitments/clinical_study_results_pharma.aspx).

However, in contrast to Adam Jacobs, I did a comprehensive investigation of this issue: I've written an R script that harvests all trials reported on Sanofi's site, and checks them against the master data file of the TrialsTracker project (by filtering all.csv to those trials that were sponsored by Sanofi and are overdue).

Let me make one thing clear: what Sanofi is doing represents a very bad practice in my opinion. (And frankly, I have no idea on why they're not uploading the results to clinicaltrials.gov. It means minimal work; I can't even think of malicious reasons for not doing this...) But, and in that I agree with Adam Jacobs, it is also unfair to call these trials "unreported". They're badly reported, sure, but not unreported.

According to my results, there are 285 Sanofi trials in TrialTracker's database that is listed as "overdue", and from them, 227 (79.6%) can be found on the above page!

In other words, amongst the negatives for Sanofi (minimally) 79.6% means false negative! Unfortunately this pretty much invalidates TrialsTracker's findings (about Sanofi, of course) in my opinion.

This situation may be true for other drug companies, I did not have time yet to investigate this issue.

Of course, for complete picture, we should not forget that not only false negatives, but false positives might arise due to TrialsTracker "automated" method. So, to have a fair picture, those that are reported in TrialsTracker as non-overdue should also be more rigorously checked, because mistakes in them might lead to the opposite error, i.e. the overestimation of the reporting rate.

To be clear, I very much like automated methods like that of TrialsTracker, but this underlines the importance of validation, and – if necessary – the fine-tuning of those automated algorithms. For transparency, the R code I used for the above investigation:

```
library( rvest )
all <- read.csv( "all.csv", stringsAsFactors = FALSE )
SanofiOverdue <- subset( all, lead_sponsor=="Sanofi"&is_overdue=="True" )$nct_id
url <-
"http://en.sanofi.com/Innovation/clinical_trials/our_commitments/clinical_study_results_pharma.aspx"
SanofiReported <- do.call( rbind, html_table( html_nodes( read_html( url ), "table" ), fill = TRUE ) ) [ , 6 ]
table( SanofiOverdue%in%SanofiReported )
```

Competing Interests: Non

Reader Comment 07 Nov 2016

Adam Jacobs, Associate Director, Biostatistics, Premier Research, UK

Powell-Smith and Goldacre have clearly put an impressive amount of work into developing an automated tool to determine the extent of undisclosed trials. However, if their tool cannot accurately determine the extent of undisclosed trials, then those efforts have been unsuccessful.

What would make this paper more convincing would be if the sensitivity and specificity of their method were to be calculated by comparison against a gold standard of a thorough manual search. It does not seem that Powell-Smith and Goldacre have done this. Although they have done what they describe themselves as "rudimentary checks" of the validity of their data, there is no calculation of specificity and sensitivity, and the checks are based on a sample of limited scope. It is not clear why that sample was chosen, and whether it was prospectively chosen or chosen post-hoc.

I was curious to see how well their method performed, so I downloaded their raw data and looked up the first 10 "undisclosed" trials sponsored by Sanofi, as this was the sponsor with the largest number of "undisclosed" trials according to the Trials Tracker website. Those trials had the trial identifiers NCT00069888, NCT00081796, NCT00087802, NCT00087958, NCT00094081, NCT00094965, NCT00103649, NCT00104013, NCT00115570, and NCT00123565.

All except 2 of those trials had their results disclosed on [Sanofi's own website](#). Presumably Powell-Smith and Goldacre's algorithm missed them as it did not check any sources except clinicaltrials.gov and Pubmed, so would miss sponsor websites. Of the remaining 2, one (NCT00094081) was [published in a peer-reviewed journal](#) (but without the publication mentioning the clinical trials ID, so it would also be missed by an automated search), and only one (NCT00123565) remained undisclosed after a 5 minute search of Google and Pubmed. Trial NCT00123565 was of a drug which was abandoned in clinical development in 2008, so no patient is deprived of information on a drug they are taking by the failure to disclose that study.

I do not know whether those 10 trials I happened to pick are representative. However, if they are, it suggests that Powell-Smith and Goldacre have overestimated the number of undisclosed trials by a factor of 10. This would make their results useless for any practical purpose.

Competing Interests: I have previously written articles (such as <http://www.statsguy.co.uk/zombie-statistics-on-half-of-all-clinical-trials-unpublished/>) criticising the All Trials campaign for exaggerating the extent of unpublished trials.
