Finding Success in Clinical Trial Recruitment: A Trial Registry Analysis

CURRENT STATUS: UNDER REVIEW

Alex Butler
Columbia University

Ziran Li
Columbia University

Cong Liu
Columbia University

Casey Ta
Columbia University

Chi Yuan
Columbia University

Huanyao Chen
Columbia University

Bruce Levin
Columbia University

George Hripcsak
Columbia University

Chunhua Weng cw2384@cumc.columbia.edu

Corresponding Author
ORCiD: 0000-0002-9624-0214

DOI:
10.21203/rs.2.19499/v1

SUBJECT AREAS
Integrative & Complementary Medicine Internal Medicine

KEYWORDS
Clinical Trials, Clinical Research, Patient Recruitment, Trial Design, Trial Registry Analysis
Abstract

Background: Randomized clinical trials are the gold-standard for generating high-quality medical evidence, but patient recruitment remains one of the most important barriers to their success despite significant administrative effort and money being spent to address this problem. While previous studies have highlighted key trial design characteristics, such as trial phase, trial sponsor, and high target accrual, as important factors in why some trials fail to recruit enough patients, these studies have been limited in the number of trials analyzed and in the scope of trial characteristics considered. This work aims to thoroughly assess the association of different trial characteristics on patient enrollment in terms of recruitment rate and early termination rate on a larger scale than has been accomplished previously.

Methods: This trial registration analysis collected recruitment information on clinical trials registered in ClinicalTrials.gov as well as trial characteristics from multiple additional databases (Clinical Trials Transformation Initiative, COHD.io, automatically parsed eligibility criteria). Descriptive statistics were calculated and the primary outcomes were associations of individual trial characteristics with patient recruitment rate and likelihood of early termination due to failed patient recruitment as well as variable selection using Group LASSO.

Results: The trial characteristics with the strongest significant associations to patient recruitment included design variables (e.g. intervention model, allocation status, number of locations, phase, etc.), sponsor experience (e.g. sponsor class, number of previous trials terminated due to recruitment issues, ratio of terminated trials to completed trials, etc.), eligibility criteria (e.g. number of inclusion criteria, number of exclusion criteria), and trial competition (e.g. overlapping eligibility criteria, similar trials within 100 miles, etc.). Different disease categories also showed different recruitment efficacy.
Conclusions: When designing clinical trials, special attention should be paid to design variables, sponsor trial experience, eligibility criteria, and trial competition to balance the likelihood of successful recruitment against evidence strength. Further research is needed to identify causal variables and improve the predictive power of patient recruitment rates to increase the breadth of this analysis.

Background

Randomized clinical trials (RCTs) are the gold standard for generating high-quality medical evidence (1). The validity of RCTs depends on successful enrollment (1, 2), which remains the number one barrier to the success of RCTs. According to recent statistics, only 2–4% of adult patients with cancer participate in RCTs, and this number has relatively remained unchanged since 1994 (2, 3). Inefficient or unrepresentative participant recruitment can cause study delays, increase costs, weaken the statistical power of analysis, and finally, lead to failed clinical trials (4).

A major bottleneck in RCT recruitment is eligibility screening (2). Recent studies have shown that up to 85% of clinical trials fail to retain enough patients (including recruitment and retention), leading to almost 4 out of every 5 trials failing to finish on time despite nearly $1.9 billion being spent on recruitment in 2012 alone (2). Many trials are also forced to extend their recruitment periods to boost recruitment, further increasing trial cost (5). As such, careful planning of clinical trials is of paramount importance in ensuring the trials continue to completion. Pharmaceutical companies and medical institutions often conduct power analyses and sample size assessments when designing clinical trials, but despite these efforts, less than one third of publicly funded trials manage to recruit patients according to their original plan (6–8).

Due to this patient recruitment dilemma, a number of analyses have assessed the impact of individual clinical trial characteristics on patient recruitment in an effort to understand
how best to structure and design new studies (9–11). Tang et al. focused on oncology trials at MD Anderson Cancer Center and found that sponsor type, longer time to first enrollment, and high target accrual were associated with slow recruitment rate (12). For clarity, the official National Cancer Institute definition for trial sponsor is “a person, company, institution, group, or organization that oversees or pays for a clinical trial and collects and analyzes the data” (13). Bennette et al. used cancer trial characteristics from ClinicalTrials.gov (CT.gov), along with manually-curated recruitment information, to further identify this research (14), but the data collected were fairly limited and only focused on cancer trials, restricting the generalizability of their findings. This study extends these previous works and includes information about the trials’ target disease(s), sponsor historical trial information, and eligibility criteria to better characterize recruitment success. We hypothesize that the relative population representativeness of eligibility criteria, as well as past trial experience of investigators and sponsors, will be strongly associated with the recruitment rate or early termination rate of clinical trials. Further, we hypothesize the size of eligible population, estimated using Electronic Health Record (EHR) data, will be associated with trial recruitment. By investigating the above hypotheses, we aim to provide additional information for future study developers to consider when designing new clinical trials.

Methods

Trial Selection and Curation

As of February 2019, 15,602 clinical trials were extracted from ClinicalTrials.gov (CT.gov) which provided recruitment details in the ‘participant flow’ table of the Aggregate Analysis of ClinicalTrials.gov (AACT) database (15). To improve sample homogeneity, trials were further filtered by interventional in the study_type field, treatment in the primary_purpose field, no in the healthy_volunteers field and actual in the enrollment_type field suggesting
the trial’s recruitment phase had ended. Total patient enrollment was extracted from the enrollment field and the dates of first and last enrollment were extracted from free text using SUTime from the Stanford NLP Group (16). A manual review of extracted data was also performed to ensure accuracy.

Data Extraction and Annotation

To enable robust aggregate analysis, this dataset was further annotated. Due to inconsistencies in trial records, information about the 8 trial design parameters outlined in Table 1 were sought in both structured data fields (e.g. phase, start date, etc.) and free text data fields (e.g. official title, brief summary, etc.). The keyword-based searches used in the free text data fields (11) were developed following a manual analysis of a subset of 300 randomly selected clinical trials.

Design Information

Specifically, phase was classified into 5 levels based on the value in the phase field:

‘NA’, if the value is empty or N/A
‘phase 0/1’, if the value is Early Phase 1 or Phase 1
‘phase 2’, if the value is Phase 1/Phase 2 or Phase 2
‘phase 3’, if the value is Phase 2/Phase 3 or Phase 3
‘phase 4’, if the value is Phase 4

Center status was categorized as ‘single-center’ or ‘multi-center’ according to the number of locations provided in facilities table.

Control status was filled with ‘controlled’ if:

1. the trial has an arm of type Placebo comparator, No Intervention, Sham comparator, or Active comparator AND
2. the trial has an intervention whose name includes the phrase placebo, vehicle, or sugar pill AND
3. the trial’s brief title, official title or brief summary contains any of the following phrases: controlled, active-controlled, active comparator, comparative study,
non-inferiority, standard therapy, standard of care, or standard treatment

Otherwise, the control status was filled with ‘non-controlled.’

Agency Class was mapped to three sponsor classes:

‘NIH/US Fed’, if NIH or US Fed was listed either as the lead sponsor of a trial or a collaborator
‘Industry’, if NIH was not involved, but either the lead sponsor or a collaborator was from industry
Remaining trials were assigned ‘Other Investigators’

The following variables were classified based on the values in corresponding fields in

AACT: intervention models (‘parallel assignment’ vs. ‘factorial assignment’ vs. ‘single
group assignment’ vs. ‘crossover assignment’), allocation status (‘randomized’ vs. ‘non-
randomized’), masking status (‘blind’ vs. ‘open label’), and data monitoring committee
(DMC) status (‘has dmc’ vs. ‘no dmc’). A criteria entity is defined as the concept of the
Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)

Vocabulary (17) presented in the eligibility criteria of the clinical trial.

Wherever possible, missing information was inferred from available fields. For example,
for single arm interventional trials with missing allocation and masking status, they were
assigned as ‘non-randomized’ and ‘open label’; for trials with missing location data, the
center status was inferred from the information in countries table, overall_officials table,
official_title field and recruitment_details field.

Recruitment Success

The dataset was further enriched with derived variables. We derived recruitment time
from dates of first and last enrollment. Each trial’s average recruitment rate was
calculated by the total number of patients recruited divided by recruitment time in units
of participants per week.

Additionally, each clinical trial in this dataset was further labeled as ‘active’, ‘completed’,
‘terminated due to recruitment issue’, or ‘terminated due to other issues’ based on the
information in overall_status and why_stopped fields (specific criteria outlined below in
sponsor trial experience). Trials which were active but no longer recruiting at the time of analysis were considered ‘completed’ as they had completed their patient recruitment. For the sake of this analysis, trials stopped early due to non-recruitment related reasons were excluded, leaving 3,077 trials for analysis hereby referred to as the target trials. A flowchart of these trial selection steps is shown in Fig. 1.

Sponsor Trial Experience

All study sponsors with at least one occurrence of lead in the lead_or_collaborator field were identified. For each sponsor, the count of previous trials as well as count of trials according to overall status (completed, terminated, actively recruiting, etc.) were collected. Further, trials were considered ‘stopped’ if they had withdrawn, suspended or terminated in the overall_status field and were considered ‘terminated due to recruitment issue’ if the following terms were found in the why_stopped field: enroll, recruit, accrue, eligibility, eligible, lack of patient, not enough patient, no patient. The following additional characteristics were calculated using these counts:

- Percent completed ($\frac{\text{# completed trials}}{\text{# total trials}} \times 100$; $pct\_completed$)
- Percent stopped ($\frac{\text{# stopped trials}}{\text{# total trials}} \times 100$; $pct\_stopped$)
- Percent stopped due to recruitment failure ($\frac{\text{# stopped trials due to recruitment issue}}{\text{# total trials}} \times 100$; $pct\_stopped\_recruitment$)
- Ratio stopped vs. completed ($\frac{\text{# stopped trials}}{\text{# completed trials}}$; $ratio\_stopped\_to\_completed$)

Sponsor Specialization

In addition to trial counts, sponsor specialization was calculated as the highest percentage of trials within a single disease category (e.g. 66% of clinical trials within the Neoplasm disease category) as defined by MeSH terms, described in greater detail below.

Disease Categories

The mesh_term field was extracted from the browse_conditions table within AACT for all available clinical trials. These MeSH terms were then matched to MeSH codes according to the 2019 MeSH files in ASCII format (18). As trials can be associated with more than one
MeSH code, MeSH categories were defined for each trial by extracting the first three characters of each associated MeSH code.

Nearby Trial Locations

A dictionary of all dates where at least one target trial was recruiting patients was generated (9/30/1991–6/30/2017) and all interventional trials active on each date, according to CT.gov registration information, was tallied. Due to the limited accessibility of date information within CT.gov, trials were considered active between the start_date and primary_completion_date in the studies table. Then, for each target trial, the total number of unique clinical trials which had overlapping active periods of at least one day, and shared a MeSH target disease category, was counted. Finally, using zip_codes in the facilities table and the pyzipcode v1.0 python package, the number of active trials within 100 miles of each target trial site was calculated and titled trials within 100 miles.

Because this package only uses zip codes within the US, the scope of nearby trials was limited to sites within the United States.

Eligibility Criteria Statistics

Leveraging the sizable amount of eligibility criteria information stored within CT.gov, previous work has been done to parse and map these eligibility criteria to medical concepts within the OMOP CDM (17). Using Criteria2Query’s pre-parsed eligibility criteria, total inclusion and exclusion criteria counts were extracted. Additionally, as these criteria were mapped to the OMOP common data model, counts of inclusion and exclusion criteria were also extracted according to OMOP concept type (e.g. measurement, procedure, condition).

Individual Eligibility Criteria Analysis and Retrieval

Individual medical concepts used in eligibility criteria were extracted for each trial and were simplified using concept clustering as described previously (19). For each medical
concept, the total count as an eligibility criterion across all trials in ClinicalTrials.gov was calculated and was titled concept count. Additionally, for each medical concept and each target trial, the count of times a concept is used as an eligibility criterion in related trials according to MeSH category (e.g. in other autoimmune diseases) was calculated and titled overlap count.

Thirdly, for each medical concept, a competition score was calculated using a formula modified from Liu et al.: (19)

\[
\sum_i \left( \frac{p'_i}{p} \right) \times \log \left( \frac{p'_i}{p} \right)
\]

where \( p \) is the total count and \( p'_i \) is the concept count for overlapping disease category \( i \).

The higher the competition score means more competition a trial might face during recruitment from other trials. These data were then averaged for each target trial for all medical concepts in the eligibility criteria.

Concept Prevalence

The prevalence of each medical concept in a real patient population was extracted from Columbia University’s Observational Health Data Sciences and Informatics (OHDSI) (20) OMOP database. Records were limited to data available between 2013 and 2017, and was retrieved using Columbia Open Health Data’s (COHD’s) open-source API (http://cohd.io/api) (21). These values were then averaged for all medical concepts used as inclusion and exclusion criteria in each target trial (i.e. concept prevalence [inclusion]).

EHR Eligible Patient Population

For each target trial, recruitment rate was calculated by the number of enrolled patients divided by recruitment time (participants/week). We also calculated the recruitment rate for a subset of 255 trials conducted within New York-Presbyterian/Columbia University
Irving Medical Center (NYP/CUIMC) using the actual patients recruited within NYP/CUIMC. An EHR eligible size (i.e., the number of patients in the EHR estimated to be eligible for the trial) was calculated for each trial in the subset as the number of unique patients returned after querying the NYP/CUIMC clinical warehouse based on the major eligibility criteria (manually selected by one of the co-authors) within the recruitment period. An EHR eligible rate was calculated as eligible size divided by recruitment time (patients/week).

Statistical Analysis

For association analysis with recruitment rate, a one-way ANOVA test was conducted for each categorical variable and linear regression was conducted for each continuous variable. For association analysis with termination status, logistic regression was conducted for each continuous variable, and a chi-squared test was conducted for each categorical variable. For both recruitment rate and termination rate, 95% confidence intervals were calculated for each level of categorical variables and each tercile of continuous variables. Additionally, for clarity of results, trial characteristics with correlation of ≥ 0.50 to other trial characteristics and without clear clinical difference were excluded from this analysis.

To further explore the dependency and impact of the design parameters and sponsor type, 1000 bootstrap iterations of Group LASSO analysis (22) were conducted for both recruitment rate and early trial termination. In each bootstrap dataset, hyper-parameters were optimized based on a five-fold cross validation and a group LASSO model with optimized hyper-parameters was then fitted using this full bootstrap dataset for variable selection. Variables were selected for inclusion when the magnitude of their coefficients were larger than zero at least 95% of all bootstrap datasets. Subsequently, multivariate linear and logistic regressions were fitted using the selected variables for recruitment rate
and early termination, respectively. In addition, linear regression was performed on the NYP/CUIMC subset to test the association between EHR eligible rate and recruitment rate. All statistical analyses were performed using R (3.3.3) (23) and Group LASSO was performed using the gglasso package (24). Significance was determined as p-value < 0.05.

Results

Descriptive Statistics

Trials included in this analysis were active from 9/30/1991 to 6/30/2017. Among the 3,077 trials, 2,789 trials (90.6%) were ‘completed’ and 288 trials (9.4%) were ‘stopped due to recruitment failure.’ The average recruitment rate was 3.71 participants per week. Descriptive statistics are available in Fig. 2 and Tables 1–4.

Design Factors

Many design factors show a significant association with both recruitment rate and termination rate (Table 5 Design factors). Compared to single and cross-over design, factorial and parallel design show a higher recruitment rate and lower termination rate. In trials with multiple centers, recruitment rate is increased, and trials are less likely to terminate early due to recruitment issues. Blind trials have a higher recruitment rate and lower termination rate than open labeled trials. Controlled trials have a higher recruitment rate and lower termination rate than non-controlled trials. DMC status does not show a significant effect on recruitment process. Phase 3 trials have a significantly higher recruitment rate than all other phases, and Phase 3 and Phase 4 trials have a lower termination rate than Phase 1 and Phase 2 trials.

Focusing on design parameters, Control Status and DMC status were excluded by Group LASSO relating to both recruitment rate and early trial termination with Intervention Model and Sponsor Class also being removed relating to early trial termination. Multivariate logistic regression showed a significant positive association of open label masking and
single center design with Early Trial Termination ($p < 0.001$; Table 6). Multivariate linear regression showed a significant positive association of Phase 3 trial (comparing to unknown phases), blind masking, and multiple center design, and industry sponsor with Patient Recruitment Rate ($p < 0.001$; Table 6).

Eligibility Criteria

In general, fewer eligibility criteria in trial protocols are associated with higher recruitment rate and decreased termination rate. This trend is statistically significant for inclusion eligibility criteria. Trials with fewer than 3 inclusion eligibility criteria in the protocol have an average recruitment rate of 5.52 patients/week, which is about 3 times higher than trials with more than 8 inclusion criteria (Table 5 Number of Criteria).

Sponsors

Trials sponsored by industry have a significantly higher recruitment rate and lower termination rate than those sponsored by the National Institute of Health. Both sponsor’s specialization and sponsor’s trial termination history were strongly associated with slower patient recruitment and increased likelihood of early termination (Table 5 Sponsors).

Competing Trials

More competing trials targeting the same group of eligible patients, represented by higher overlap counts (Table 5 Competing Trials), are shown to have lower recruitment rate and higher termination rate. In addition, the number of nearby trials was shown to have a significant association with recruitment rate and early trial termination. Patient prevalence of concepts used in eligibility criteria also showed significant association with recruitment rate and trial termination, though this differed between inclusion and exclusion criteria.

Target Disease

Different target conditions and interventions can have different recruitment rates and
termination rates. Nutritional and metabolic diseases have the highest recruitment rate, followed by cardiovascular diseases and pathological conditions (Fig. 3).

EHR Eligible Rate
We further investigated whether the number of eligible patients in EHR is associated with the recruitment process using 255 trials conducted with NYP/CUIMC. Linear regression identified a significant association between the patient recruitment rate and number of eligible patients in the EHR with a slope coefficient of $9.82 \times 10^{-5}$ and standard error of $2.45 \times 10^{-5}$ (p-value < 0.001; Fig. 4).

Discussion
In this study, in an effort to further explore how observational data and historical clinical trial experience impact patient recruitment in clinical trials, we identified trial characteristics including design factors, sponsors, competing trials and eligibility criteria with significant association with both recruitment rate and early trial termination. Further, using a subset of trials conducted in NYP/CUIMC and its EHR data, we explored the correlation between the size of the eligible patient population and the trial recruitment process. This study provides new depth and clarity into the association between clinical trial design and participant enrollment.

Patient recruitment has been reported to vary widely across different trial designs sponsors, conclusions which are supported here (25–29). As previous research suggests, trials led by pharmaceutical companies with the aim of releasing a new drug or identifying a new indication are more likely to receive high levels of funding and administrative input (27), possibly explaining this elevated recruitment success (25, 26). To account for this complex relationship, careful variable selection was conducted using a Group LASSO analysis and the remaining important factors were sponsor type, phase, masking status
and number of recruiting centers, highlighting their impact on patient recruitment. Significant associations were also identified among trial sponsor experience, or more specifically, the sponsor’s historical completed and stopped trials and disease specialization. Patient recruitment requires significant financial and administrative investment (30) and it could thus be assumed that more experienced sponsors with a broad clinical focus are more capable of meeting these demands. However, some sponsors specialize in rare diseases in an effort to combat their historic difficulties with finding eligible patients (25, 31). In a separate Group LASSO analysis (results not shown here), disease specialization was excluded, but historical experience remained, suggesting that some difficulties recruiting patients may relate to the previous success of the sponsor itself (30). To the best of our knowledge, this is the first study to show that a sponsor’s historical experience is strongly correlated to its future performance, regardless of sponsor class, and should inspire future research into this area in an effort to establish a causal relationship.

Though not a new idea, quantitative evidence was provided here that eligibility criteria are also an important factor in patient recruitment with increased numbers of inclusion criteria limiting patient recruitment and increasing the likelihood of early termination (2, 32, 33). However, the same association was not present for the number of exclusion criteria. This conclusion was further supported by the differing association of concept prevalence within inclusion and exclusion criteria in relation to early trial termination. The roles of inclusion and exclusion criteria are reported to be different with inclusion criteria including demographic, clinical, and geographical characteristics and exclusion criteria defining features which could interfere with the success of the study or increase the risk of an unfavorable outcome (34). As exclusion criteria are responsible for limiting the study population to those who have the greatest chance to experience clinical benefit and avoid
unnecessary risk, these criteria are also understood to hamper the trials generalizability (35–37). Further, these criteria are often not justified by their authors, leading to a lack of uniform application across trials targeting the same disease or drug (37, 38). As such, their inconsistent use among related trials is one possible explanation for their relatively low impact on patient recruitment in this analysis, though further experimentation is required to confirm and expand this hypothesis.

Beyond the number of criteria alone, trial competition also showed a strong association with patient recruitment. This is a relatively well recognized phenomenon with the Clinical Trials Transformation Initiative (CTTI) proposing a series of actionable recommendations toward improving patient recruitment in 2018 which included site selection based on access to the target population (8). In this work, however, we were able to quantify this negative association with patient recruitment and provided new means to quantitatively assess the competition, helping to inform future protocol planning and site selection methods moving forward (14, 39).

EHRs have been universally adopted in almost every hospital, clinic, and other healthcare institution for collecting medical data of patients. EHRs have previously been suggested as a key way to improve the likelihood of completing trial accrual in a timely fashion (40) and have been shown to improve efficiency in E-screening for research coordinators (41). Results presented here further show there is a clear association between the size of institution-wide eligible population and trial recruitment efficacy with an increase in 0.05 patients recruited per week for every 500 additional eligible patients. As evidenced here, EHR data has the potential to optimize trial recruitment prediction in a single clinical site though further research is required to highlight its clinical utility (42).

As mentioned in the introduction, the primary aim of this study was to assess the relationship between certain clinical trial characteristics and patient recruitment. Though
not all of these characteristics can be easily adjusted to improve recruitment, such as phase, sponsor, or targeted disease, many characteristics highlighted here serve as clear guidelines on how investigators can improve their chances of successful patient enrollment. For example, when designing eligibility criteria, increased attention should be paid to reducing restrictiveness to allow for optimal patient recruitment. Complex criteria are often defended on the grounds of safety concerns, but recent studies have shown that as few as 47.2% of criteria are justified in any particular RCT (37). This implies that relaxation of eligibility restrictiveness can be pursued without exposing patients to unnecessary risk. In fact, a recent analysis tool, titled GIST, was developed just for this purpose (37, 43). Further, improved access to clinical site information should be leveraged to carefully choose where new sites are to open, avoiding direct competition with nearby trials and allowing patients in more areas to have access to these options.

LIMITATIONS OF THIS STUDY

This study has several limitations. First, because of the nature of retrospective design, our analysis was unable to establish causality between the collected variables and successful patient recruitment. Another limitation is that automated eligibility criteria parsing can be incomplete and occasionally inaccurate (17). In this work, many pitfalls were avoided by limiting the analysis to eligibility metadata (e.g. total counts of inclusion and exclusion criteria, averaged patient prevalence) and not including individual criteria or concepts. However, this approach also serves to limit the scope of trial characteristics assessed. Finally, though this study expanded the scope of characteristic study performed previously, this was not a complete list. Much of the information regarding clinical trials remains in free-text within CT.gov, making access over a large number of clinical trials very difficult and error-prone as the complexity has been well described (9). Future work in this field should include more longitudinal data collection, improved automated natural
language processing, and a greater expanse of trial information to address these stated limitations and further improve our understanding of patient recruitment.

Conclusions

Although patient recruitment is a well-recognized barrier to clinical trial success, little was understood about the role of specific trial characteristics on this issue. In this analysis, multiple categories of clinical trial characteristics were found to be strongly associated with patient recruitment and trial success, including design parameters, eligibility criteria restrictiveness/competition and sponsor trial experience, and quantitative support was provided to support previous hypotheses. As increased success of clinical trials can mean greater and faster availability of novel therapies for patients, adopting a more informed approach to trial design may provide a new way for investigators to give studies the best chance of success.

Abbreviations

AACT
Aggregate Analysis of ClinicalTrials.gov
ANOVA
Analysis of Variance
CDM
Common Data Model
COHD
Columbia Open Health Data
CT.gov
ClinicalTrials.gov
CTTI
Clinical Trials Transformation Initiative
CUIMC
Columbia University Irving Medical Center
DMC
Declarations

**Ethics Approval and Consent to Participate**

Not applicable as this study did not involve the use of any animal or human data or tissue.

**Consent for Publication**

Not applicable as this study does not contain any data from any individual person.

**Availability of Data and Materials**

In addition to data sources listed and cited in the Methods section above, the datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

**Competing Interests**

The authors declare that they have no competing interests.

**Funding**

This research was supported by National Library of Medicine Research grants
R01LM009886, UL1TR001873, and Janssen Pharmaceuticals (grant reference CU15-2317). The study’s design, collection, analysis, and interpretation were independent of these funding sources as well as the writing of the report and decision to submit the article for publication.

Authors’ Contributions

These three authors contributed equally in the primary research and authorship role: AB, ZL, CL. CT, CY, HC assisted with study design and data extraction. BL and GH assisted with statistical analysis and data interpretation. CW managed the project and serves as primary guarantor. All authors contributed to drafting and revising of the final manuscript.

Acknowledgements

Not applicable.

References

1. Gul RB, Ali PA. Clinical trials: the challenge of recruitment and retention of participants. J Clin Nurs. 2010;19(1-2):227-33.

2. Penberthy LT, Dahman BA, Petkov VI, DeShazo JP. Effort Required in Eligibility Screening for Clinical Trials. Journal of Oncology Practice. 2012;8(6):365-70.

3. Joseph RR. Viewpoints and concerns of a clinical trial participant. Cancer. 1994;74(S9):2692-3.

4. Campillo-Gimenez B, Buscail C, Zekri O, Laguerre B, Le Prisé E, De Crevoisier R, et al. Improving the pre-screening of eligible patients in order to increase enrollment in cancer clinical trials. Trials. 2015;16.

5. Moffat KR, Cannon P, Shi W, Sullivan F. Factors associated with recruitment to randomised controlled trials in general practice: protocol for a systematic review. Trials. 2019;20(1):266.

6. Barnard KD, Dent L, Cook A. A systematic review of models to predict recruitment to
multicentre clinical trials. BMC Medical Research Methodology. 2010;10(1):63.

7. Institute NC. Report of the Operational Efficiency Working Group of the Clinical Trials in Translational Research Advisory Committee: Compressing the Timeline for Cancer Clinical Trial Activation. 2010.

8. Huang GD, Bull J, Johnston McKee K, Mahon E, Harper B, Roberts JN. Clinical trials recruitment planning: A proposed framework from the Clinical Trials Transformation Initiative. Contemporary Clinical Trials. 2018;66:74-9.

9. Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. Summit Transl Bioinform. 2010;2010:46-50.

10. Robinson L, Adair P, Coffey M, Harris R, Burnside G. Identifying the participant characteristics that predict recruitment and retention of participants to randomised controlled trials involving children: a systematic review. Trials. 2016;17(1):294.

11. Zwierzyna M, Davies M, Hingorani AD, Hunter J. Clinical trial design and dissemination: comprehensive analysis of clinicaltrials.gov and PubMed data since 2005. Bmj. 2018;361:k2130.

12. Tang C, Sherman SI, Price M, Weng J, Davis SE, Hong DS, et al. Clinical Trial Characteristics and Barriers to Participant Accrual: The MD Anderson Cancer Center Experience over 30 years, a Historical Foundation for Trial Improvement. Clin Cancer Res. 2017;23(6):1414-21.

13. Institute NC. NCI Dictionary of Cancer Terms.18.12e.

14. Bennette CS, Ramsey SD, Mc Dermott CL, Carlson JJ, Basu A, Veenstra DL. Predicting Low Accrual in the National Cancer Institute’s Cooperative Group Clinical Trials. J Natl Cancer Inst. 2015;108(2):djv324.

15. Aggregate Analysis of ClinicalTrials.gov. Clinical Trials Transformation Initiative.

16. Chang A, Manning C. SUTIME: A Library for Recognizing and Normalizing Time
Expressions2012.

17. Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. Journal of the American Medical Informatics Association: JAMIA. 2019;26(4):294-305.

18. Medicine NLo. Medical Subject Headings (MeSH). 2019.

19. Liu C, Yuan C, Butler AM, Carvajal RD, Li ZR, Ta CN, et al. DQueST: dynamic questionnaire for search of clinical trials. Journal of the American Medical Informatics Association. 2019.

20. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-8.

21. Ta CN, Dumontier M, Hripcsak G, Tatonetti NP, Weng C. Columbia Open Health Data, clinical concept prevalence and co-occurrence from electronic health records. Scientific Data. 2018;5:180273.

22. Gertheiss J, Tutz G. Sparse modeling of categorial explanatory variables. Ann Appl Stat. 2010;4(4):2150-80.

23. Team RC. R: A language and environment for statistical computing. Vienna, Australia: R Foundation for Statistical Computing; 2019.

24. Yang Y, Zou H. gglasso: Group Lasso Penalized Learning Using a Unified BMD Algorithm. 2017.

25. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. Contemporary clinical trials communications. 2018;11:156-64.

26. Haddad RI, Chan AT, Vermorken JB. Barriers to clinical trial recruitment in head and neck cancer. Oral oncology. 2015;51(3):203-11.
27. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nature Biotechnology. 2014;32:40.

28. Pinato DJ, Stavraka C, Flynn MJ, Forster MD, O'Cathail SM, Seckl MJ, et al. An inflammation based score can optimize the selection of patients with advanced cancer considered for early phase clinical trials. PloS one. 2014;9(1):e83279.

29. Weber JS, Levit LA, Adamson PC, Bruinooge S, Burris HA, Carducci MA, et al. American Society of Clinical Oncology Policy Statement Update: The Critical Role of Phase I Trials in Cancer Research and Treatment. Journal of Clinical Oncology. 2014;33(3):278-84.

30. Probstfield JL, Frye RL. Strategies for Recruitment and Retention of Participants in Clinical Trials. JAMA. 2011;306(16):1798-9.

31. Augustine EF, Adams HR, Mink JW. Clinical Trials in Rare Disease: Challenges and Opportunities. Journal of Child Neurology. 2013;28(9):1142-50.

32. George SL. Reducing patient eligibility criteria in cancer clinical trials. Journal of Clinical Oncology. 1996;14(4):1364-70.

33. Carlisle B, Kimmelman J, Ramsay T, MacKinnon N. Unsuccessful trial accrual and human subjects protections: an empirical analysis of recently closed trials. Clinical trials (London, England). 2015;12(1):77-83.

34. Patino CM, Ferreira JC. Inclusion and exclusion criteria in research studies: definitions and why they matter. J Bras Pneumol. 2018;44(2):84-.

35. Bugeja G, Kumar A, Banerjee AK. Exclusion of elderly people from clinical research: a descriptive study of published reports. BMJ. 1997;315(7115):1059.

36. Lee PY, Alexander KP, Hammill BG, Pasquali SK, Peterson ED. Representation of elderly persons and women in published randomized trials of acute coronary syndromes. Jama. 2001;286(6):708-13.
37. Van Spall HG, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. Jama. 2007;297(11):1233-40.

38. Schmidt AF, Groenwold RH, van Delden JJ, van der Does Y, Klungel OH, Roes KC, et al. Justification of exclusion criteria was underreported in a review of cardiovascular trials. J Clin Epidemiol. 2014;67(6):635-44.

39. Kaplan CP, Nápoles AM, Dohan D, Shelley Hwang E, Melisko M, Nickleach D, et al. Clinical trial discussion, referral, and recruitment: physician, patient, and system factors. Cancer Causes & Control. 2013;24(5):979-88.

40. Mayo CS, Matuszak MM, Schipper MJ, Jolly S, Hayman JA, Ten Haken RK. Big Data in Designing Clinical Trials: Opportunities and Challenges. Front Oncol. 2017;7:187.

41. Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D. Electronic screening improves efficiency in clinical trial recruitment. Journal of the American Medical Informatics Association : JAMIA. 2009;16(6):869-73.

42. Milo Rasouly H, Wynn J, Marasa M, Reingold R, Chatterjee D, Kapoor S, et al. Evaluation of the cost and effectiveness of diverse recruitment methods for a genetic screening study. Genet Med. 2019;21(10):2371-80.

43. Sen A, Chakrabarti S, Goldstein A, Wang S, Ryan PB, Weng C. GIST 2.0: A scalable multi-trait metric for quantifying population representativeness of individual clinical studies. J Biomed Inform. 2016;63:325-36.

Tables

**Table 1. Descriptive Statistics of Design Parameters**
| Design Parameter            | Number (%) | n = 3,077 |
|----------------------------|------------|-----------|
| Phase                      |            |           |
| Phase 0/1                   | 278 (9%)   |           |
| Phase 2                     | 1188 (39%) |           |
| Phase 3                     | 734 (24%)  |           |
| Phase 4                     | 395 (13%)  |           |
| Center Status               |            |           |
| Single-Center               | 1320 (43%) |           |
| Domestic Multi-Center       | 1117 (36%) |           |
| International Multi-Center  | 640 (21%)  |           |
| Control Status              |            |           |
| Controlled                  | 1800 (59%) |           |
| Non-Controlled              | 1277 (42%) |           |
| Allocation Status           |            |           |
| Randomized                  | 1826 (59%) |           |
| Non-Randomized              | 1251 (41%) |           |
| Intervention Model          |            |           |
| Single Group Assignment     | 1155 (38%) |           |
| Parallel Assignment         | 1757 (57%) |           |
| Factorial Assignment        | 26 (1%)    |           |
| Crossover Assignment        | 139 (5%)   |           |
| Masking Status              |            |           |
| None (Open Label)           | 1863 (61%) |           |
| Single                      | 199 (6%)   |           |
| Double                      | 331 (11%)  |           |
| Triple                      | 233 (8%)   |           |
| Quadruple                   | 451 (15%)  |           |
| Sponsor Class               |            |           |
| Industry                    | 1607 (52%) |           |
| Investigator                | 694 (23%)  |           |
| NIH/US Federal              | 776 (25%)  |           |
| DMC Status                  |            |           |
| Has DMC                     | 1392 (45%) |           |
| No DMC                      | 1685 (55%) |           |

Table 2. Descriptive Statistics of Target Conditions
| Target Condition Category [MeSH Category] | Number (%), n = 3,077 |
|------------------------------------------|-----------------------|
| Neoplasms [C04]                          | 1150 (37%)            |
| Pathological Conditions, Signs and Symptoms [C23] | 699 (23%)        |
| Immune System Diseases [C20]              | 438 (14%)             |
| Cardiovascular Diseases [C14]             | 397 (13%)             |
| Hemic and Lymphatic Diseases [C15]        | 349 (11%)             |
| Digestive System Diseases [C06]           | 311 (10%)             |
| Nervous System Diseases [C10]             | 292 (9%)              |
| Skin and Connective Tissue Diseases [C17] | 282 (9%)              |
| Nutritional and Metabolic Diseases [C18]  | 267 (9%)              |
| Mental Disorders [F03]                    | 267 (9%)              |
| Female Urogenital Diseases and Pregnancy Complications [C13] | 266 (9%)   |
| Endocrine System Diseases [C19]          | 256 (8%)              |
| Respiratory Tract Diseases [C08]         | 245 (8%)              |
| Male Urogenital Diseases [C12]           | 232 (8%)              |
| Musculoskeletal Diseases [C05]           | 144 (5%)              |
| Virus Diseases [C02]                      | 129 (4%)              |
| Congenital, Hereditary, and Neonatal Diseases and Abnormalities [C16] | 99 (3%)        |
| Behavior and Behavior Mechanisms [F01]   | 96 (3%)               |
| Bacterial Infections and Mycoses [C01]   | 85 (3%)               |
| Eye Diseases [C11]                       | 82 (3%)               |
| Chemically-Induced Disorders [C25]       | 57 (2%)               |
| Wounds and Injuries [C26]                | 49 (2%)               |
| Stomatognatic Diseases [C07]             | 43 (1%)               |
| Otorhinolaryngologic Diseases [C09]      | 39 (1%)               |
| Physiological Phenomena [G07]            | 30 (1%)               |
| Diagnosis [E01]                          | 22 (1%)               |
| Psychological Phenomena [F02]            | 13 (0%)               |
| Musculoskeletal and Neural Physiological Phenomena [G11] | 11 (0%)   |
| Parasitic Diseases [C03]                 | 7 (0%)                |
| Investigative Techniques [E05]           | 6 (0%)                |
| Tissues [A10]                            | 3 (0%)                |
| Reproductive and Urinary Physiological Phenomena [G08] | 3 (0%)          |
| Environment and Public Health [N06]      | 3 (0%)                |
| Cells [A11]                              | 2 (0%)                |
| Physical Phenomena [G01]                 | 2 (0%)                |
| Cell Physiological Phenomena [G04]       | 2 (0%)                |
| Biological Phenomena [G16]               | 2 (0%)                |
| Social Sciences [I01]                    | 2 (0%)                |
| Animal Diseases [C22]                    | 1 (0%)                |
| Occupational Diseases [C24]              | 1 (0%)                |
| Genetic Phenomena [G05]                  | 1 (0%)                |

Table 3. Descriptive Statistics of Additional Trial Characteristics
Table 4. Descriptive Statistics of Trial Outcomes

| Outcome                        | Mean (Std) | Number (%), n = 3,077 |
|--------------------------------|------------|------------------------|
| Recruitment Rate               | Rate per Week | 3.74 (11.81)           |
| Total Trial Timeline           | Length in Weeks | 117.78 (93.1)         |
| Overall Status                 | Terminated due to Recruitment Failure | 288 (9%) |
|                                | Completed/Active, Not Recruiting     | 2,789 (91%) |

Table 5: Association analysis of factors affect recruitment rate and termination rate
Table 6: LASSO Analysis and Multivariate Regression Analysis of Design Parameters
| Trial Parameters | Early Trial Termination | Recruitment Rate (pts/week) |
|------------------|-------------------------|-----------------------------|
|                  | LASSO Excluded | Estimate | Std. Error | p-value | LASSO Excluded | Estimate | Std. Error | p-value |
| Phase 0/1        | 2.1%          | 0.32     | 0.34       | 0.35    | 0%           | -0.13    | 0.17       | 0.43    |
| Phase 2          | 2.1%          | 0.69     | 0.28       | 0.14    | 0%           | -0.09    | 0.13       | 0.49    |
| Phase 3          | 2.1%          | 0.26     | 0.34       | 0.45    | 0%           | 0.84     | 0.15       | < 0.001 |
| Phase 4          | 2.1%          | 0.08     | 0.37       | 0.83    | 0%           | 0.13     | 0.16       | 0.42    |
| Control Status*  | 28.1%         | à        | à          | à       | 56.8%        | à        | à          | à       |
| Masking Status*  | 0%            | 0.65     | 0.20       | < 0.001 | 2.9%         | -0.26    | 0.10       | 0.009   |
| Center Status*   | 0%            | 0.81     | 0.17       | < 0.001 | 0%           | -0.33    | 0.09       | < 0.001 |
| Intervention Model* | 21.9% | à        | à          | à       | 4.2%         | -0.19    | 0.10       | 0.067   |
| DMC Status*      | 21.1%         | à        | à          | à       | 97.8%        | à        | à          | à       |
| Sponsor Class*   | 5.8%          | à        | à          | à       | 0%           | -0.41    | 0.06       | <0.001  |

* Regression estimate for Control Status refers to Non-Controlled, Masking Status refers to Open-Label, Center Status refers to Single-Center, Intervention Model refers to Single/Crossover Assignment, DMC Status refers to True, Sponsor Class refers to NIH/Investigator

à These trial parameters were excluded by Group LASSO

**Figures**

298,172 trials in ClinicalTrials.gov as of February 1, 2019

17,591 trials with information in the participant_flows table

16,672 trials with study_type of Interventional
12,279 trials with primary_purpose of Treatment

11,495 trials with healthy_volunteers of No

11,494 trials with enrollment_type of Actual

3,356 trials with valid recruitment information following manual review

3,077 trials which were not stopped due to non-recruitment-related reasons
Figure 1
1. Trial Selection Flowchart a. Count of analyzed trials at each step in the trial selection process.

![Trial Selection Flowchart](image)

Figure 2
2. Distribution of Categorical Design Parameters a. Bar plot depicting frequency of trials falling into each of the categorical design parameters. Parameters grouped by category

![Categorical Design Parameters](image)
3. Target Disease Categories and Recruitment Rate

a. Target disease categories are plotted against patient recruitment rate (measured in patients per week). For clarity, only categories within the DISEASES [C] and PSYCHIATRY AND PSYCHOLOGY [F] MeSH subheadings are shown in this plot.
4. Association Between the Eligible Population and Recruitment Rate a. X-axis is the number of unique patients (normalized by week) returned after querying the NYP/CUIMC clinical warehouse based on the major eligibility criteria within the recruitment period. Y-axis is the number of patients enrolled in NYP/CUIMC (normalized by week) for the corresponding target trial.

Supplementary Files

This is a list of supplementary files associated with the primary manuscript. Click to download.

butler_strobe_checklist.docx